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

Weiming Zhang[1] Yexin Liu[1] Xu Zheng[1] Lin Wang[1,2*]
[1] AI Thrust, HKUST(GZ) [2]Dept. of CSE, HKUST
zweiming996@gmail.com, yliu292@connect.hkust-gz.edu.cn, zhengxu128@gmail.com, linwang@ust.hk
Project Page: https://vlislab22.github.io/GoodSAM/

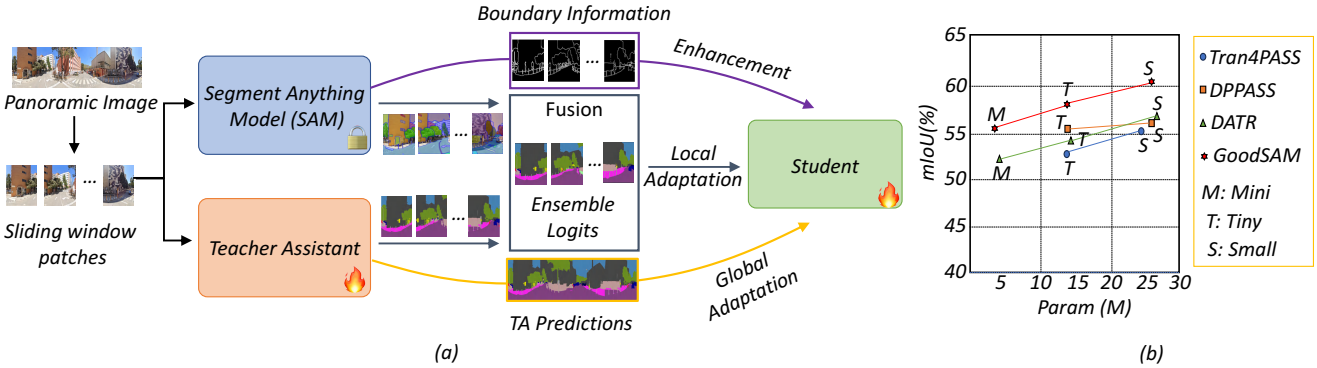

Figure 1. (a) Illustration of our GoodSAM, leveraging instance masks and boundary information provided by SAM, coupled with segmentation logits from the teacher assistant, to obtain reliable ensemble logits for knowledge adaptation to our student. (b) Our GoodSAM outperforms SOTA methods [37, 41, 42] across various model parameter ranges. Notably, GoodSAM-M achieves comparable performance to the SOTA methods with only **3.7**M parameters.

Abstract

This paper tackles a novel yet challenging problem: how to transfer knowledge from the emerging Segment Anything Model (SAM) – which reveals impressive zero-shot instance segmentation capacity – to learn a compact panoramic semantic segmentation model, i.e., *student, without requiring any labeled data. This poses considerable challenges due to SAM's inability to provide semantic labels and the large capacity gap between SAM and the student. To this end, we propose a novel framework, called* ***GoodSAM****, that introduces a teacher assistant (TA) to provide semantic information, integrated with SAM to generate ensemble logits to achieve knowledge transfer. Specifically, we propose a Distortion-Aware Rectification (DAR) module that first addresses the distortion problem of panoramic images by imposing prediction-level consistency and boundary enhancement. This subtly enhances TA's prediction capacity on panoramic images. DAR then incorporates a cross-task complementary fusion block to adaptively merge the predictions of SAM and TA to obtain more reliable ensemble logits. Moreover, we introduce a Multi-level Knowledge Adaptation (MKA) module to efficiently transfer the multi-level feature knowledge from TA and ensemble logits to learn a compact student model. Extensive experiments on two benchmarks show that our GoodSAM achieves a remarkable* ***+3.75%*** *mIoU improvement over the state-of-the-art (SOTA) domain adaptation methods,* e.g., *[41]. Also, our most lightweight model achieves comparable performance to the SOTA methods with only* ***3.7****M parameters.*

1. Introduction

The burgeoning demand for omnidirectional scene understanding has stimulated the popularity of 360° cameras, making them highly suitable and popular for applications, such as self-driving [6, 21, 22, 41]. This has inspired intensive recent research endeavors [8, 36, 42] for addressing scene understanding tasks, especially panoramic semantic segmentation. Generally, Equirectangular Projection (ERP) is commonly adopted to represent the 360° data in 2D planar representation [1] to align spherical data with the existing

*Corresponding Author

[1]In this paper, omnidirectional and panoramic images are interchangeably used, and ERP images often indicate panoramic images.

deep learning pipelines. However, ERP images often suffer from distortion and deformation problems, induced by the non-uniformly distributed pixels[1]. Also, there is a scarcity of well-annotated datasets, which poses a challenge in training effective models for panoramic segmentation.

Therefore, research has been conducted to transfer knowledge from the labeled pinhole image domain to the unlabeled panoramic image domain via unsupervised domain adaptation (UDA) [9, 36, 37, 41, 42]. These methods can be primarily divided into three groups: pseudo labeling [9, 34], adversarial training [11, 42] and prototypical adaptation [37, 38]. However, they require labeled 2D images for training or adopting multi-branch designs [29, 42], thus leading to considerable computational costs. Recently, foundational models have undergone a significant development [7, 14, 20]. Large visual models (LVMs), such as the Segment Anything Model (SAM) [7] – trained on very large datasets (over 1 billion masks on 11 million images) have received great attention. The exceptional zero-shot instance segmentation performance on unseen datasets and tasks makes SAM exceptional in serving as a foundational model for many segmentation tasks [12, 23, 39].

In this paper, we tackle a novel problem: *how to transfer the instance segmentation knowledge from SAM to learn a more compact panoramic semantic segmentation model (*i.e.*, student) without requiring any labeled data?* This poses considerable challenges to directly apply SAM for panoramic semantic segmentation due to two reasons: **1)** SAM's inability to provide semantic labels, and **2)** the substantial capacity gap between SAM and the student model. These obstacles render the task of learning a compact student model non-trivial. To this end, we propose a novel framework, named **GoodSAM** to learn a compact student model. Our key insight is to introduce a teacher assistant (TA) to 1) provide semantic labels, enabling to generate the ensemble logits with SAM, and 2) bridge the capacity gap with the student for effective knowledge adaptation.

Our GoodSAM enjoys two key technical contributions. Specifically, we first propose a Distortion-Aware Rectification (DAR) module to generate reliable ensemble logits (Sec. 3.1). DAR module optimizes ensemble logits from three aspects, see Fig. 1 (a). Firstly, we leverage an overlapping sliding window strategy [10] to alleviate the adverse effects of the ERP image's large Field of View (FoV) on the performance of both SAM and TA. Secondly, we impose prediction-level consistency for overlapping regions between two adjacent windows and boundary enhancement, using the boundary information provided by SAM, to enhance TA's ability to address the inevitable distortion and object deformation in the ERP images. Finally, we propose a cross-task complementary fusion (CTCF) block, which adaptively combines SAM's instance masks and TA's semantic labels to obtain *high-quality ensemble logits*. Upon obtaining reliable and distortion-aware ensemble logits, we then introduce the Multi-level Knowledge Adaptation (MKA) module to facilitate the learning of a compact panoramic segmentation student model (Sec. 3.2). MKA facilitates multi-level and multi-scale knowledge transfer from TA and ensemble logits, encompassing both whole-image scale and window-based scale to bridge the capacity gap between SAM and the student model and improve the performance of our compact student model.

We conducted extensive experiments to validate our method. As shown in Fig. 1 (b), our GoodSAM outperforms SOTA UDA methods across various model parameter ranges. Our GoodSAM's small version, with approximately **25** million parameters, achieves an impressive **3.75%** performance improvement compared to SOTA methods with a similar parameter count. Meanwhile, our GoodSAM's tiny version achieves comparable performance to the SOTA methods while using only **3.7** million parameters.

In summary, our contributions are as follows: **(I)** Our work serves as the **first** attempt to learn an efficient panoramic semantic segmentation model from SAM. **(II)** We propose the GoodSAM framework which incorporates DAR and MKA modules to obtain reliable ensemble logits and conduct effective knowledge transfer for panoramic segmentation, respectively. **(III)** We demonstrate the effectiveness of our proposed GoodSAM framework, achieving SOTA performance on panoramic semantic segmentation tasks while maintaining a compact model size.

2. Related Work

Panorama Image Semantic Segmentation. The first line of works [15, 25, 27–30] on panoramic semantic segmentation are based on the supervised learning. However, since there is no sufficient panoramic image datasets exist, most of the existing panoramic image semantic segmentation methods are based on unsupervised domain adaptation (UDA) [44, 45]. Recent research endeavors have been focused on the UDA for panoramic semantic segmentation approaches, which can be divided into three types, including the pseudo labeling [9, 34, 40], adversarial training [11, 36, 42] and prototypical adaptation [37, 38, 43] methods. *Differently, we introduce SAM to the panoramic semantic segmentation task, aiming at transferring the instance segmentation knowledge of SAM to learn a compact student model, assisted by a TA model.*

Segment Anything Model (SAM) It stands as a foundation model for segmentation tasks [7]. SAM leverages a dataset of 11 million diverse, high-resolution images, accompanied by 1.1 billion meticulously annotated high-quality segmentation masks for training. This remarkable dataset empowers SAM with robust zero-shot instance segmentation capacity. SAM has been applied to diverse domains, such as medical image segmentation [12, 18, 23, 39], image editing

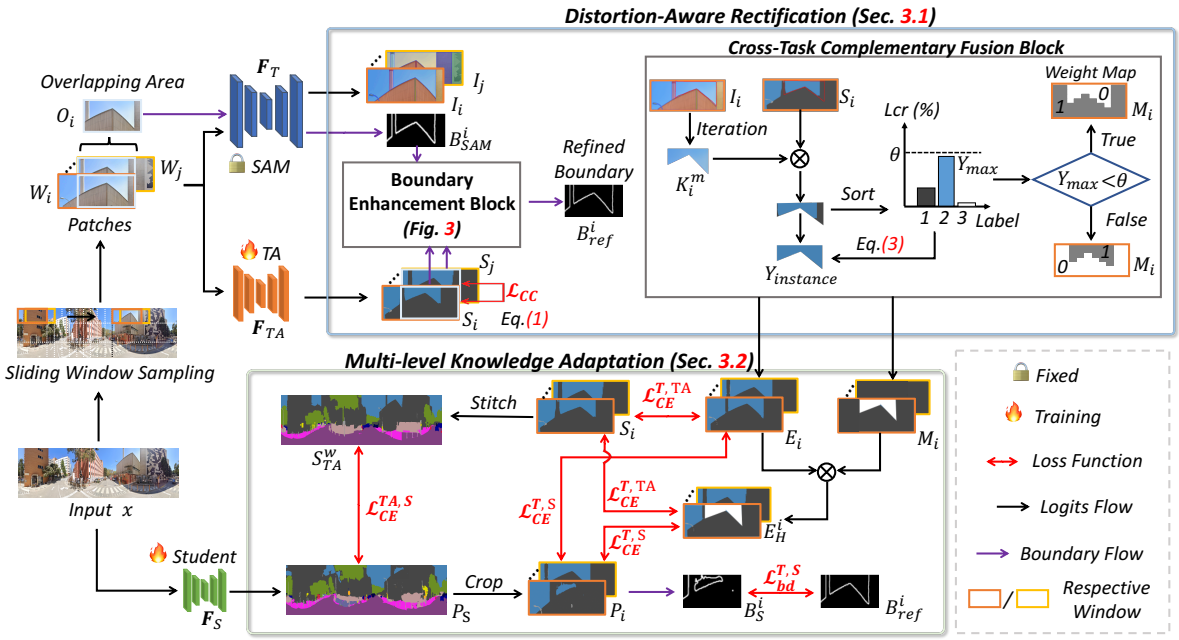

Figure 2. **Overview of GoodSAM framework**, consisting of three models: SAM, teacher assistant, and student. Our method has two main technical components: the Distortion-Aware Rectification (DAR) module and the Multi-level Knowledge Adaptation (MKA) module.

[32, 35], and tracking [4, 26]. However, SAM is limited in providing semantic information, coupled with its prohibitive parameters and domain gap. This poses a challenge for directly achieving panorama semantic segmentation. To address this, we introduce a TA model that leverages SAM's high-quality instance masks to generate reliable ensemble logits, enabling us to learn a compact yet effective panoramic segmentation model.

Knowledge Adaptation. Some methods leverage the expertise of multiple teachers and the similarities between their domains to improve the performance of a student model [5, 13, 19]. The knowledge adaptation lies in identifying the trustworthy knowledge of the teacher's expertise that can be applied effectively in the target domain. However, existing methods exhibit limitations, such as the risk of the student excessively relying on a teacher's biases, and challenges in generalizing knowledge across tasks or unseen data. Consequently, directly applying these methods to our panoramic segmentation task is impractical. Intuitively, we propose CTCF block to adaptively combine the outputs from SAM and TA to obtain more reliable ensemble logits. Meanwhile, we propose the MKA module that utilizes TA's outputs and th ensemble logits from the DAR module to perform multi-level and multi-scale knowledge transfer.

3. Methodology

Overview. An overview of our framework is shown in Fig. 2. Given an unlabeled panoramic image $x \in R^{H\times W\times 3}$, we aim to train a compact panoramic semantic segmentation model (*i.e.*, student) $\mathbf{F}_S$ under the guidance of SAM $\mathbf{F}_T$ (*i.e.*, teacher) and the assistance of the teacher assistant (TA) $\mathbf{F}_{TA}$. Note that the TA aims to bridge the capacity gap between SAM and the student during training. To mitigate the impact of ERP's large FoV, we employ an overlapping sliding window strategy to extract n local patches $\{W_i\}_{i=0}^{n-1}$ from the input ERP images. Since the distortion in horizontal pixel distribution is more severe than the vertical one, we choose to do horizontal window sliding. By setting the sliding window movement stride, we can obtain the overlapping area O_i between two horizontally adjacent windows W_i and W_j. Subsequently, patches are input to both SAM $\mathbf{F}_T$ and TA $\mathbf{F}_{TA}$, yielding their respective predictions I_i and S_i. Furthermore, the overlapping area O_i is exclusively input to $\mathbf{F}_T$ to derive the corresponding boundary map (B^i_{SAM}). For the student model $\mathbf{F}_S$, we input the ERP image x to obtain the semantic prediction map P_S. The challenges lie in: 1) effectively fusing the predictions from SAM $\mathbf{F}_T$ and TA $\mathbf{F}_{TA}$ to obtain more reliable ensemble logits E_i as the supervision for student model $\mathbf{F}_S$; 2) effectively performing knowledge adaptation from ensemble logits E_i and $\mathbf{F}_{TA}$ to our compact student $\mathbf{F}_S$. To this end, we introduce the GoodSAM framework consisting of two key technical modules: Distortion-Aware Rectification (DAR) Module (Sec. 3.1) and Multi-level Knowledge Adaptation (MKA) Module (Sec. 3.2). We now describe these modules in detail.

3.1. Distortion-Aware Rectification (DAR) Module

DAR module aims to generate ensemble logits based on SAM and TA. Specifically, as ERP images, in contrast to 2D images, possess a large FoV and distortion problem, impacting the performance of SAM and TA. We introduce the

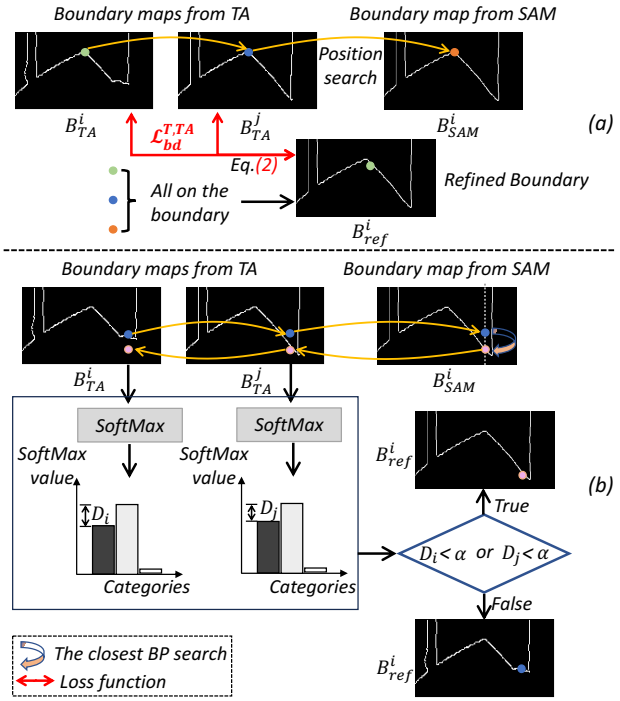

Figure 3. **Overview of the proposed boundary enhancement block**. In (a), it represents the condition where the pixels at the same positions in all three images are boundary pixels. In (b), it represents the condition where pixels at the same positions are not on the boundary. Additionally, (a) demonstrates the optimization of the boundary enhancement loss for B^i_{TA} and B^j_{TA}.

consistency constraint and boundary enhancement. Moreover, SAM and TA generate different kinds of segmentation maps: SAM provides instance masks, while TA produces semantic maps. Therefore, we propose the cross-task complementary fusion block. We now illustrate the details.

Consistency Constraint. It aims to help TA generate distortion-aware semantic maps. For the adjacent patches W_i and W_j, due to distortion of ERP, $\mathbf{F}_{TA}$'s predictions S_i and S_j for the overlapping area O_i between W_i and W_j may exhibit discrepancies. Therefore, the purpose of the constraint is to minimize the discrepancies in the overlapping area. For simplicity, we utilize the mean squared error (MSE) loss to ensure the consistency of predictions. As such, we can enhance the $\mathbf{F}_{TA}$'s sensitivity to local distortions. Formally, the consistency constraint loss $\mathcal{L}_{CC}$ is:

$$\mathcal{L}_{CC} = MSE(W_i(O_i), W_j(O_j)), \tag{1}$$

where the $W_i(O_i)$ denotes the TA's prediction within the overlapping area for the i th window.

Boundary Enhancement Block. Due to SAM's strong zero-shot capability to provide relatively accurate boundary maps, we propose a boundary enhancement block *to refine boundary pixels in TA's predictions*, inspired by [17]. By improving TA's accuracy in predicting boundary pixels, we aim to alleviate the impact of distortion and object deformation on TA. This block comprises two components: boundary refinement strategy and boundary-enhanced loss.

As depicted in Fig. 3(a), for two adjacent windows W_i and W_j, we obtain two separate boundary maps B^i_{TA} and B^j_{TA} of the overlapping area O_i. The boundary refinement strategy is proposed to identify reliable boundary pixels (denoted as green, blue and orange points) within the overlapping area by combining B^i_{TA}, B^j_{TA}, and B^i_{SAM} of SAM to obtain the refined boundary map B^i_{ref}. The detailed Algorithm for this strategy can be found in the *Supplmat*.

Specifically, for the input boundary map B^i_{TA}, we first iterate through its boundary pixels and find corresponding pixels in B^j_{TA} and B^i_{SAM} at the same positions. If the corresponding pixels at the same positions are all on the boundary (See Fig. 3 (a)), then the pixel is considered a reliable boundary pixel. For cases that do not meet the above condition, we find corresponding pixels in B^i_{SAM} at the same positions and identify the nearest boundary pixel (the pink point) in the vertical direction (See Fig. 3 (b)).

Subsequently, we locate pixels at the same positions (pink points) in B^i_{TA} and B^j_{TA}, and for each pixel, perform softmax on its logits ($1 \times 1 \times C$, where C is the number of categories). We calculate the difference in the top two softmax values for each corresponding pixel in B^i_{TA} and B^j_{TA}, denoting them as D_i and D_j respectively. When either D_i or D_j has a value less than α, we determine that the boundary pixel from B^i_{SAM} exhibits the characteristics of the boundary pixel in the other two boundary maps as well. Thus, we define the SAM boundary pixel (pink point) as reliable. The parameter α determines the influence strength of SAM boundary pixels. Finally, if none of the above conditions are met, we decide to retain the boundary pixels of B^i_{TA} as reliable pixels. This way, we attain a refined boundary map B^i_{ref} with all reliable boundary pixels for O_i, which is utilized for updating TA $\mathbf{F}_{TA}$ and student $\mathbf{F}_S$.

Next, we introduce a boundary-enhanced loss (See Fig. 3 (a)) to encourage TA's boundary pixel predictions to align closely with the refined boundary map:

$$\mathcal{L}^{T,TA}_{bd} = \sum_{k=1}^{H\times W} \frac{(|B^i_{ref} - B^i_{TA}| + |B^i_{ref} - B^j_{TA}|)}{C_o}, \tag{2}$$

where C_o denotes the total boundary pixel counts of B_{ref} and k denotes the k-th pixel in the boundary map. This explicitly mitigates boundary pixel prediction errors caused by the distortion of ERP.

Cross-Task Complementary Fusion (CTCF). To obtain more reliable ensemble logits E_i for window-based regions, we propose the CTCF block, as shown in Fig. 2. It adaptively fuses SAM $\mathbf{F}_T$'s instance mask outputs I_i with TA $\mathbf{F}_{TA}$'s semantic segmentation outputs S_i. The objective of the fusion is to assign the highest-confidence semantic label to each instance mask based on the logits S_i from TA (See Fig. 2). Different from directly finding the most frequent or area-dominant semantic label [2, 3], we define dis-

tinct area proportion thresholds for masks of different sizes and adaptively select the most reliable label. Specifically, for each instance mask K_i^m from SAM, we assess the instance mask area. For larger and smaller instance masks, we set the threshold θ to a smaller value to facilitate obtaining the most frequently occurring label. For medium-sized masks, we increase θ to a larger value to ensure the acquisition of a more accurate semantic label. Then, we identify the top three semantic labels in descending order of quantity within the corresponding area in TA's predictions. If the label coverage rate *(lcr)* of the most prevalent semantic label Y_{max} exceeds the θ, we directly assign the most prevalent label as the semantic label for the current instance mask K_i^m. If the coverage rate of the most prevalent label falls below θ, we delve into the three semantic labels and calculate their Shannon entropy (SE) using the logits S_i from TA. The label with the minimum entropy is chosen as the highest-confidence semantic label $Y_{instance}$ for the K_i^m. The formulation is as follows:

$$Y_{instance} = \begin{cases} Y_{max}, & lcr(Y_{max}) \geq \theta, \\ Y_{argmin\{SE(Y_a)\}}, & 0 < lcr(Y_a) < \theta, \end{cases} \tag{3}$$

where a belongs to the top three semantic labels occupying the instance mask. The detailed pseudo-code can be found in the *Supplmat*. Through the CTCF block, DAR produces high-quality ensemble logits by adaptively merging the predictions based on windows from both SAM and TA.

Due to the fusion process potentially resulting in varying prediction confidences among masks, we obtain a weight map (M_i) with spatial dimensions identical to W_i in the fusion process. Specifically, we assign higher weight (1) to masks with higher overlap between instance masks and semantic logits and lower weight (0) to masks that require SE for label assignment. The ensemble logits and weight maps obtained by the CTCF block can assist the TA and the student in achieving better supervision for patch predictions.

3.2. Multi-level Knowledge Adaptation Module

After addressing the distortion problem with our DAR module, we then propose the MKA module to learn a compact student module with TA $\mathbf{F}_{TA}$'s output logits S_i and the ensemble logits E_i from the DAR module for multi-level, multi-scale (whole-image and patch) knowledge adaptation. To effectively transfer knowledge from the whole-image scale prediction of the TA $\mathbf{F}_{TA}$ to the student model $\mathbf{F}_S$, we concatenate predictions of non-overlapping patches to generate the entire ERP semantic prediction map S_{TA}^w. Therefore, when the entire image is directly fed to the student model $\mathbf{F}_S$, resulting in the prediction map P_S^w. We use the Cross-Entropy (CE) loss (Eq.(5)) to guide the student in aligning its predictions with S_{TA}^w at whole-image scale:.

$$\mathcal{L}_{ce}^{TA,S} = \mathcal{L}_{CE}(S_{TA}^w, P_S^w). \tag{4}$$

For the obtained ensemble logits E_i based on the patch W_i, we use another CE loss to guide the student's prediction logits at the corresponding window position P_S^i to mimic E_i. As the CTCF block returns the weight maps M_i corresponding to the higher confidence masks based on the fusion mechanism, we first combine weight maps with ensemble logits to obtain higher confidence masks E_H^i. Then we perform knowledge adaptation simultaneously using ensemble logits E_i and higher confidence masks E_H^i. Therefore, the loss for patches knowledge from E_i and E_H^i transfer to the student $\mathbf{F}_S$ and TA $\mathbf{F}_{TA}$ can be formulated as:

$$\mathcal{L}_{CE}^{T,S} = \sum_{k=1}^{H\times W} (\mathcal{L}_{CE}(P_i^k, E_i^k) + \lambda M_i \mathcal{L}_{CE}(P_i^k, E_i^k)), \tag{5}$$

$$\mathcal{L}_{CE}^{T,TA} = \sum_{k=1}^{H\times W} (\mathcal{L}_{CE}(S_i^k, E_i^k) + \lambda M_i \mathcal{L}_{CE}(S_i^k, E_i^k)), \tag{6}$$

where λ is the hyper-parameter. By considering higher confidence masks E_H^i, we can enhance the learning of both the student and TA towards the reliable ensemble logits during the update process. This helps mitigate the impact of noisy labels generated by SE in lower confidence masks on both the student and TA. The patches supervision not only assists the student in refining the segmentation of large objects but also contributes to improved recognition of smaller objects.

Additionally, we utilize the refined boundary map B_{ref}^i obtained from DAR to supervise the student's boundary map (B_S^i) in the corresponding overlapping area O_i, enhancing the student's awareness of boundaries.

$$\mathcal{L}_{bd}^{T,S} = \sum_{k=1}^{H\times W} \frac{|B_{ref}^i - B_S^i|}{C_o}. \tag{7}$$

Therefore, the total loss employed for the student $\mathbf{F}_S$ comprises three components:

$$\mathcal{L}_{student} = \mathcal{L}_{CE}^{TA,S} + \mathcal{L}_{CE}^{T,S} + \mathcal{L}_{bd}^{T,S}. \tag{8}$$

The total loss for the TA $\mathbf{F}_{TA}$ is formulated as follows:

$$\mathcal{L}_{TA} = \mathcal{L}_{CE}^{T,TA} + \mathcal{L}_{CC} + \mathcal{L}_{bd}^{T,TA}. \tag{9}$$

4. Experiments

4.1. Datasets and Implementation Details

Dataset. We leverage two benchmark datasets WildPASS [31] and DensePASS [11] to assess the segmentation performance of the GoodSAM. 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. 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 [24] series, encompassing B0-B5 variants, which come in six different sizes and exhibit varying performance levels in 2D image semantic segmentation. We set the window size as 400×512, with stride 256. The hyper-parameter

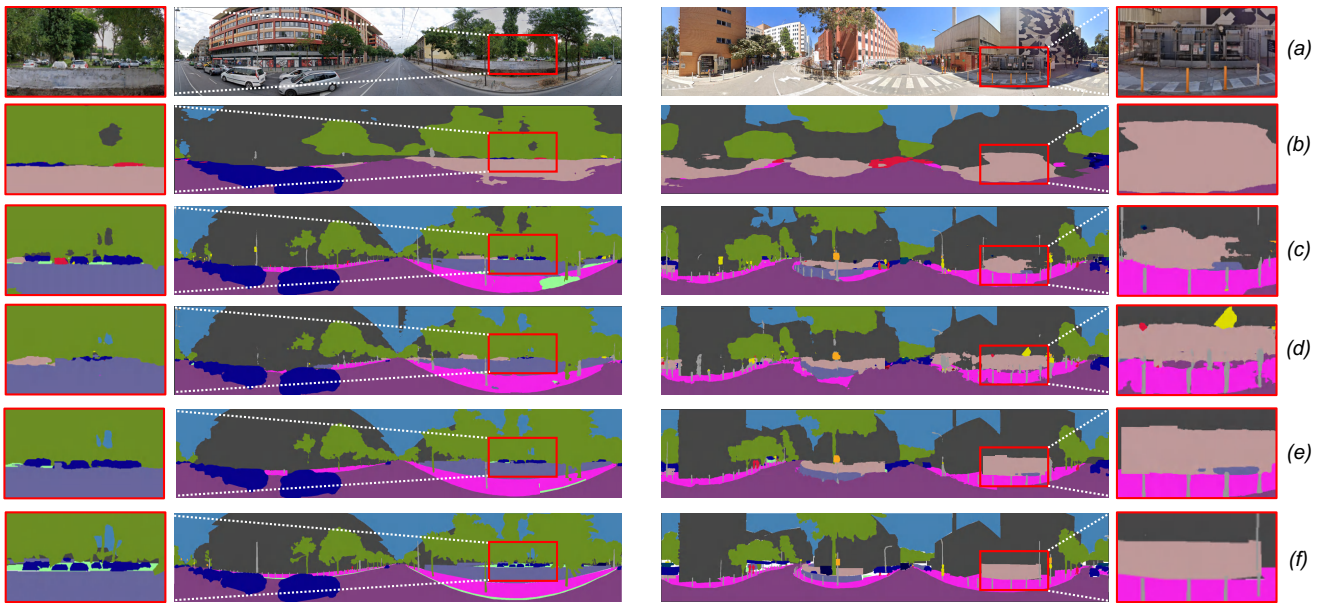

Figure 4. Example visualization results from the DensePASS test set: (a) Input panorama image, (b) Segformer-B5 [24] without sliding window sampling, (c) DPPASS-S [42], (d) DATR-S [41], (e) GoodSAM-S, (f) Ground truth.

Method	P. (M)	mIoU	Road	S.W.	Build.	Wall	Fence	Pole	Tr.L.	Tr.S.	Veget.	Terr.	Sky	Person	Rider	Car	Truck	Bus	Train	M.C.	B.C.
ERFNet [16]	-	16.65	63.59	18.22	47.01	9.45	12.79	17.00	8.12	6.41	34.24	10.15	18.43	4.96	2.31	46.03	3.19	0.59	0.00	8.30	5.55
PASS(ERFNet) [28]	-	23.66	67.84	28.75	59.69	19.96	29.41	8.26	4.54	8.07	64.96	13.75	33.50	12.87	3.17	48.26	2.17	0.82	0.29	23.76	19.46
Omni-sup(ECANet) [30]	-	43.02	81.60	19.46	81.00	32.02	39.47	25.54	3.85	17.38	79.01	39.75	94.60	46.39	12.98	81.96	49.25	28.29	0.00	55.36	29.47
P2PDA(Adversarial) [36]	-	41.99	70.21	30.24	78.44	26.72	28.44	14.02	11.67	5.79	68.54	38.20	85.97	28.14	0.00	70.36	60.49	38.90	77.80	39.85	24.02
PCS [33]	25.56	53.83	78.10	46.24	86.24	30.33	45.78	34.04	22.74	13.00	79.98	33.07	93.44	47.69	22.53	79.20	61.59	67.09	83.26	58.68	39.80
Trans4PASS-T [37]	13.95	53.18	78.13	41.19	85.93	29.88	37.02	32.54	21.59	18.94	78.67	**45.20**	93.88	48.54	16.91	79.58	65.33	55.76	84.63	59.05	37.61
Trans4PASS-S [37]	24.98	55.22	78.38	41.58	86.48	31.54	45.54	33.92	22.96	18.27	79.40	41.07	93.82	48.85	23.36	81.02	67.31	**69.53**	86.13	60.85	39.09
DPPASS-T [42]	14.0	55.30	78.74	46.29	87.47	**48.62**	40.47	35.38	24.97	17.39	79.23	40.85	93.49	52.09	29.40	79.19	58.73	47.24	86.48	66.60	38.11
DPPASS-S [42]	25.4	56.28	78.99	48.14	87.63	42.12	44.85	34.95	27.38	19.21	78.55	43.08	92.83	**55.99**	29.10	80.95	61.42	55.68	79.70	70.42	38.40
DATR-M [41]	4.64	52.90	78.71	48.43	86.92	34.92	43.90	33.43	22.39	17.15	78.55	28.38	93.72	52.08	13.24	77.92	56.73	59.53	93.98	51.52	34.06
DATR-T [41]	14.72	54.60	79.43	49.70	87.39	37.91	44.85	35.06	25.16	19.33	78.73	25.75	93.60	53.52	20.20	78.07	60.43	55.82	91.11	67.03	34.32
DATR-S [41]	25.76	56.81	80.63	51.77	87.80	44.94	43.73	**37.23**	25.66	**21.00**	78.61	26.68	93.77	54.62	**29.50**	80.03	67.35	63.75	87.67	67.57	37.10
GoodSAM-M(ours)	3.7	55.93	79.57	51.04	86.24	43.42	44.86	30.92	26.60	20.62	77.79	25.43	92.99	53.77	25.84	82.01	70.94	62.29	91.93	58.24	38.25
GoodSAM-T(ours)	14.0	58.21	80.06	**53.29**	89.75	44.91	46.98	31.13	27.81	19.83	79.58	25.72	93.81	55.44	26.99	84.54	73.07	68.41	93.99	67.36	43.39
GoodSAM-S(ours)	25.4	**60.56**	**80.98**	52.96	**93.22**	48.17	**51.28**	33.51	**28.09**	20.15	**81.64**	30.97	**95.21**	55.13	29.01	**87.89**	**75.28**	69.37	**94.98**	**73.28**	**49.64**

Table 1. Per-class results of the SOTA panoramic image semantic segmentation methods on DensePASS test set. (P.: Param.)

α is set to 0.3. For medium-sized masks, the area range 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. *More details can be found in the supplmat.*

4.2. Comparisons with Existing Works

We first compare GoodSAM with previous panoramic semantic segmentation methods, including ERFNet [16], PASS [28], Omni-sup [30], P2PDA [36], PCS [33], Trans4PASS [37], DPPASS [42], and DATR [41], on the DensePASS dataset. As shown in Tab. 1, our GoodSAM-M, GoodSAM-T, and GoodSAM-S consistently exhibit superior performance at their respective parameter levels. Specifically, Our GoodSAM-S outperforms DATR-S, DPPASS-S, and Trans4PASS-S by **3.75**% IoU, **4.28**% IoU, and **5.34**% IoU, respectively, which yields SOTA performance. Additionally, GoodSAM-M achieves a competitive mIoU of **55.93**% mIoU with just **3.7** million parameters, which is comparable to DATR-T and surpasses Trans4PASS-T. For the segmentation performance of each class, our GoodSAM-S outperforms the others in the majority of classes, including 'building' (+**5.42**% IoU), 'fence' (+**5.5**% IoU), and nearly all types of transportation (*e.g.*, 'car' with +**5.93**% IoU). Fig. 4 shows visual comparisons of GoodSAM-S with other methods on the Densepass evaluation set. This showcases that our GoodSAM-S can generate distortion-aware and boundary-enhanced logits under the supervision of SAM and TA.

Tab. 2 presents the experimental results on the WildPASS validation dataset. Our GoodSAM outperforms the existing methods based on Segformer B1 and B2 backbones. Specif-

Method	Backbone	mIoU(%)
Source domain Supervised	Segformer-B1 [24]	47.90
	Segformer-B2 [24]	54.11
Trans4PASS-T [37]	Segformer-B1	54.67
Trans4PASS-S [37]	Segformer-B2	62.91
DPPASS-T [42]	Segformer-B1	60.38
DPPASS-S [42]	Segformer-B2	63.53
GoodSAM-M(ours)	Segformer-B0	58.65
GoodSAM-T(ours)	Segformer-B1	62.42
GoodSAM-s(ours)	Segformer-B2	**65.18**

Table 2. Experimental results of the SOTA panoramic image semantic segmentation methods on WildPASS test set.

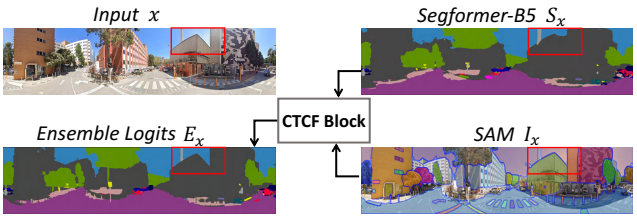

Figure 5. Effectiveness of the CTCF block.

ically, our GoodSAM-T outperforms DPPASS-T by **2.04**% mIoU on the Segformer-B1 backbone, and our GoodSAM-S surpasses DPPASS-S by **1.65**% mIoU on the Segformer-B2 backbone. This indicates that, even with an expanded number of evaluation images, our GoodSAM consistently achieves superior performance.

4.3. Ablation Studies and Analysis

4.3.1 Effectiveness of DAR module

Tab. 3 illustrates the effectiveness of each component of our framework. We choose Segformer-B5 as the TA model and Segformer-B0 as the student model.

1) Effectiveness of sliding window sampling. The performance comparison before and after employing sliding window sampling for the DAR module reveals a significant performance gap. This is attributed to Segformer being trained on 2D images, and when directly applied to ERP images with larger FoV characteristics, the model shows a substantial decline in performance. Specifically, the performance gap is nearly twice as significant without the sliding window strategy (B5: **27.62**% mIoU vs. **53.86**% mIoU, B0: **15.88**% mIoU vs. **39.35**% mIoU).

2) Effectiveness of CTCF. We evaluate the CTCF block based on the non-overlapping sliding window strategy. We propose CTCF to obtain patch ensemble logits and then rebuild the entire ERP prediction map at corresponding positions. As shown in Tab. 5, the mIoU of the ensemble logits currently stands at **55.88**% mIoU, surpassing TA's performance by **2.02**% mIoU. Furthermore, as we utilize $\mathcal{L}_{ce}^{T,TA}$ to update TA based on patch ensemble logits, we observe a continuous improvement(+**5.1**% mIoU) in ensemble logits performance alongside the enhancement of TA's performance. Fig. 5 demonstrates the visual differences between the segmentation map produced by Segformer-B5 as TA and the ensemble logits obtained through CTCF block, highlighting the effectiveness of our CTCF block.

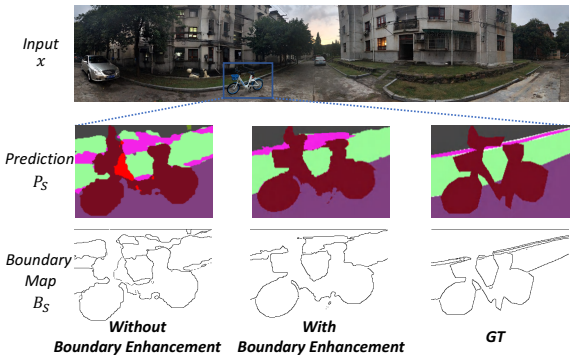

Figure 6. Effectiveness of the boundary enhancement block.

3) Effectiveness of consistency constraint. we now ablate the consistency constraint based on the overlapping sliding window. Introducing the consistency constraint $\mathcal{L}_{CC}$ to implicitly mitigate the inconsistencies in overlapping regions between adjacent windows caused by distortion significantly improves the TA's performance(+**1.18**% mIoU).

4) Effectiveness of boundary enhancement. We propose the boundary enhancement block to leverage the boundary information provided by SAM, enhancing TA's ability to predict boundary pixels. As revealed by Tab. 3, the introduced boundary-enhanced loss boosts TA's performance by **2.42**% mIoU (from **60.07**% mIoU to **62.49**% mIoU). The right part of visualization results in Fig. 4 demonstrates our superior boundary segmentation performance for 'fence' and 'sidewalk' compared to previous methods. Meanwhile, Fig. 6 also illustrates the impact of the boundary enhancement block on student predictions. These experiment results indicate that our boundary enhancement block explicitly assists TA and student in increasing awareness of boundaries and addressing distortion problem.

4.3.2 Effectiveness of MKA module

For the MKA module, as shown in Tab. 3, we observe that the performance of the student is only **15.88**% mIoU when the entire ERP image is directly input. However, as we update the student using TA's logits, the performance increases to **50.90**% mIoU with the improvement of TA's performance. Furthermore, by supervising the semantic logits of the student in the corresponding region using window-based ensemble logits, the performance of our student improves by **3.22**% mIoU. Finally, by further constraining the student's boundary pixel prediction results with the refined boundary map obtained from DAR, our student's performance has been elevated to **55.93**% mIoU. This indicates the effectiveness of MKA in endowing GoodSAM with

SAM and TA		TA			Student			mIoU		
SW	CTCF	$\mathcal{L}_{ce}^{T,TA}$	$\mathcal{L}_{CC}$	$\mathcal{L}_{bd}^{T,TA}$	$\mathcal{L}_{CE}^{TA,S}$	$\mathcal{L}_{CE}^{T,S}$	$\mathcal{L}_{bd}^{T,S}$	E.S.	TA	Student
-	-	-	-	-	-	-	-	-	27.62	15.88
√	-	-	-	-	-	-	-	-	53.86	15.88
√	√	-	-	-	-	-	-	55.88	53.86	15.88
√	√	√	-	-	√	-	-	60.68	58.89	49.88
√	√	√	√	-	√	-	-	-	60.07	50.07
√	√	√	√	√	√	-	-	-	62.49	50.90
√	√	√	√	√	√	√	-	-	62.49	54.12
√	√	√	√	√	√	√	√	-	62.49	55.93

Table 3. Ablation of loss functions during the training process. E.S. represents ensemble logits.

distortion-aware and boundary-aware capabilities.

4.3.3 Other Analysis

Hyper-parameters analysis in sliding window sampling. Tab. 4 presents the effects of four different window sizes and step sizes with Segformer-B2. We observe that TA exhibits the best performance (**53.86**%) when the window size is 400×512. Meanwhile, the ensemble logits obtained after fusion with SAM achieve the highest performance (**55.88**%). For the sliding window stride, we find that as the stride decreases, the overlapping area between adjacent windows increases, resulting in an increase in the number of windows and training overhead for TA. Therefore, based on Tab. 4, due to constraints on training resources, when the stride is set to 256, TA achieves the highest performance(**60.07**%) under the constraints of $\mathcal{L}_{CE}^{T,TA}$ and $\mathcal{L}_{CC}$.

Analysis about CTCF module. We introduce the CTCF block to adaptively combine the outputs of SAM and TA, obtaining patch ensemble logits. Now, we evaluate our CTCF by comparing it with current methods that combine instance masks and semantic logits [2, 3]. SSA [2] assigns the label of the instance mask as the label that appears most frequently in the corresponding region of the semantic map. The fusion mechanism in SEPL [3] is similar to SSA. They analyze each instance mask, selecting the label if it occupies more than half of the area in the corresponding region of the semantic map or if the distribution is almost covered by the instance mask. However, we analyze the instance masks of different sizes during the experiment. We observe that the comparable coverage rate between the top two labels often occurs in medium-sized masks, which increases the risk of errors if the label with the highest rate is directly chosen. Therefore, we set different thresholds for different area sizes and incorporate SE to identify the highest-confidence label, ensuring an adaptive fusion of instance masks and semantic logits. Tab. 5 illustrates that our fusion mechanism enhances the robustness and accuracy of the fusion process compared to the other two methods.

Hyper-parameter analysis in the loss functions. When transferring knowledge from window-based ensemble logits to the TA and student, we assign higher weights to the instance masks with high confidence for the CTCF block. Fig. 7 shows that with λ set to 0.2, our GoodSAM can more efficiently extract correct knowledge from ensemble logits.

Window Size	400×1024	400×512	400×256	400×128
mIoU	52.45	**53.86**	53.29	52.79
Strides	512	400	320	256
mIoU	58.89	59.68	59.73	**60.07**

Table 4. Ablation about threshold window size and the step size of the sliding windows.

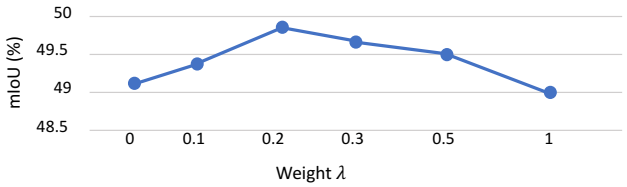

Figure 7. Ablation study for λ of knowledge tranfer from ensemble logits to our student.

		B_0	B_1	B_2	B_3	B_4	B_5
Fusion	SSA [2]	39.28	40.28	47.93	49.13	52.19	54.32
	SEPL [3]	39.68	40.75	48.98	50.37	52.53	54.58
	CTCF	40.68	41.75	49.7	52.89	54.13	55.88
Student	B_2	-	-	-	57.16	59.02	60.56

Table 5. Ablation of fusion ways and different TA model.

The impact of TA selection. We assess the impact of different TAs on the training of the student. As shown in Tab. 5, we observe that as the performance of TA improves, our student's performance also increases. Specifically, when TA is Segformer-B5, our student achieves a performance gain of **+3.4**% mIoU compared to TA being Segformer-B3. This improvement is attributed to Segformer-B5's ability to provide more comprehensive and accurate semantic logits.

5. Conclusion and Future Work

In this paper, we designed a comprehensive framework for lightweight panorama semantic segmentation, which leverages the assistance of SAM and TA. By addressing the distortion and large FoV problems in panoramic images and bridging the capacity gap between SAM and the student model, our GoodSAM produced distortion-aware and boundary-enhanced logits, surpassing SOTA UDA methods across various model parameter levels.

Future work: It would be worthwhile to fine-tune SAM to realize a foundational segmentation model suitable for panoramic images. Additionally, we plan to investigate methods for distilling SAM's zero-shot capabilities into our compact segmentation model.

Acknowledgment: This paper was supported by the Guangzhou 2024 Applied Basic Research Project (Co-funded by Municipal Schools (Institutes) and Enterprises) Fund under Grant No. 2024A04J4072.

References

[1] Hao Ai, Zidong Cao, Jinjing Zhu, Haotian Bai, Yucheng Chen, and Lin Wang. Deep learning for omnidirectional vision: A survey and new perspectives. *arXiv preprint arXiv:2205.10468*, 2022. 2

[2] Jiaqi Chen, Zeyu Yang, and Li Zhang. Semantic segment anything. https://github.com/fudan-zvg/Semantic-Segment-Anything, 2023. 4, 8

[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. 4, 8

[4] Yangming Cheng, Liulei Li, Yuanyou Xu, Xiaodi Li, Zongxin Yang, Wenguan Wang, and Yi Yang. Segment and track anything. *arXiv preprint arXiv:2305.06558*, 2023. 3

[5] Tong He, Chunhua Shen, Zhi Tian, Dong Gong, Changming Sun, and Youliang Yan. Knowledge adaptation for efficient semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 578–587, 2019. 3

[6] Maleen Jayasuriya, Ravindra Ranasinghe, and Gamini Dissanayake. Active perception for outdoor localisation with an omnidirectional camera. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4567–4574. IEEE, 2020. 1

[7] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 2

[8] Xuewei Li, Tao Wu, Zhongang Qi, Gaoang Wang, Ying Shan, and Xi Li. Sgat4pass: Spherical geometry-aware transformer for panoramic semantic segmentation. *arXiv preprint arXiv:2306.03403*, 2023. 1

[9] Mengyi Liu, Shuhui Wang, Yulan Guo, Yuan He, and Hui Xue. Pano-sfmlearner: Self-supervised multi-task learning of depth and semantics in panoramic videos. *IEEE Signal Processing Letters*, 28:832–836, 2021. 2

[10] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 2

[11] 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

[12] Jun Ma and Bo Wang. Segment anything in medical images. *arXiv preprint arXiv:2304.12306*, 2023. 2

[13] Zhigang Ma, Yi Yang, Nicu Sebe, and Alexander G Hauptmann. Knowledge adaptation with partiallyshared features for event detectionusing few exemplars. *IEEE transactions on pattern analysis and machine intelligence*, 36(9):1789–1802, 2014. 3

[14] Duy MH Nguyen, Hoang Nguyen, Nghiem T Diep, Tan N Pham, Tri Cao, Binh T Nguyen, Paul Swoboda, Nhat Ho, Shadi Albarqouni, Pengtao Xie, et al. Lvm-med: Learning large-scale self-supervised vision models for medical imaging via second-order graph matching. *arXiv preprint arXiv:2306.11925*, 2023. 2

[15] Semih Orhan and Yalin Bastanlar. Semantic segmentation of outdoor panoramic images. *Signal, Image and Video Processing*, 16(3):643–650, 2022. 2

[16] Eduardo Romera, José M Alvarez, Luis M Bergasa, and Roberto Arroyo. Erfnet: Efficient residual factorized convnet for real-time semantic segmentation. *IEEE Transactions on Intelligent Transportation Systems*, 19(1):263–272, 2017. 6

[17] Shenghai Rong, Bohai Tu, Zilei Wang, and Junjie Li. Boundary-enhanced co-training for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19574–19584, 2023. 4

[18] Saikat Roy, Tassilo Wald, Gregor Koehler, Maximilian R Rokuss, Nico Disch, Julius Holzschuh, David Zimmerer, and Klaus H Maier-Hein. Sam. md: Zero-shot medical image segmentation capabilities of the segment anything model. *arXiv preprint arXiv:2304.05396*, 2023. 2

[19] Sebastian Ruder, Parsa Ghaffari, and John G Breslin. Knowledge adaptation: Teaching to adapt. *arXiv preprint arXiv:1702.02052*, 2017. 3

[20] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15638–15650, 2022. 2

[21] Fu-En Wang, Hou-Ning Hu, Hsien-Tzu Cheng, Juan-Ting Lin, Shang-Ta Yang, Meng-Li Shih, Hung-Kuo Chu, and Min Sun. Self-supervised learning of depth and camera motion from 360 {\deg} videos. *arXiv preprint arXiv:1811.05304*, 2018. 1

[22] Fu-En Wang, Yu-Hsuan Yeh, Min Sun, Wei-Chen Chiu, and Yi-Hsuan Tsai. Bifuse: Monocular 360 depth estimation via bi-projection fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 462–471, 2020. 1

[23] Junde Wu, Rao Fu, Huihui Fang, Yuanpei Liu, Zhaowei Wang, Yanwu Xu, Yueming Jin, and Tal Arbel. Medical sam adapter: Adapting segment anything model for medical image segmentation. *arXiv preprint arXiv:2304.12620*, 2023. 2

[24] 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. 5, 6, 7

[25] Yuanyou Xu, Kaiwei Wang, Kailun Yang, Dongming Sun, and Jia Fu. Semantic segmentation of panoramic images using a synthetic dataset. In *Artificial Intelligence and Machine Learning in Defense Applications*, pages 90–104. SPIE, 2019. 2

[26] Jinyu Yang, Mingqi Gao, Zhe Li, Shang Gao, Fangjing Wang, and Feng Zheng. Track anything: Segment anything meets videos. *arXiv preprint arXiv:2304.11968*, 2023. 3

[27] Kailun Yang, Xinxin Hu, Luis M Bergasa, Eduardo Romera, Xiao Huang, Dongming Sun, and Kaiwei Wang. Can we pass beyond the field of view? panoramic annular semantic segmentation for real-world surrounding perception. In *2019 IEEE Intelligent Vehicles Symposium (IV)*, pages 446–453. IEEE, 2019. 2

[28] Kailun Yang, Xinxin Hu, Luis M Bergasa, Eduardo Romera, and Kaiwei Wang. Pass: Panoramic annular semantic segmentation. *IEEE Transactions on Intelligent Transportation Systems*, 21(10):4171–4185, 2019. 6

[29] Kailun Yang, Xinxin Hu, Hao Chen, Kaite Xiang, Kaiwei Wang, and Rainer Stiefelhagen. Ds-pass: Detail-sensitive panoramic annular semantic segmentation through swaftnet for surrounding sensing. In *2020 IEEE Intelligent Vehicles Symposium (IV)*, pages 457–464. IEEE, 2020. 2

[30] Kailun Yang, Xinxin Hu, Yicheng Fang, Kaiwei Wang, and Rainer Stiefelhagen. Omnisupervised omnidirectional semantic segmentation. *IEEE Transactions on Intelligent Transportation Systems*, 23(2):1184–1199, 2020. 2, 6

[31] 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. 5

[32] Tao Yu, Runseng Feng, Ruoyu Feng, Jinming Liu, Xin Jin, Wenjun Zeng, and Zhibo Chen. Inpaint anything: Segment anything meets image inpainting. *arXiv preprint arXiv:2304.06790*, 2023. 3

[33] Xiangyu Yue, Zangwei Zheng, Shanghang Zhang, Yang Gao, Trevor Darrell, Kurt Keutzer, and Alberto Sangiovanni Vincentelli. Prototypical cross-domain self-supervised learning for few-shot unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13834–13844, 2021. 6

[34] Cheng Zhang, Zhaopeng Cui, Cai Chen, Shuaicheng Liu, Bing Zeng, Hujun Bao, and Yinda Zhang. Deeppanocontext: Panoramic 3d scene understanding with holistic scene context graph and relation-based optimization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12632–12641, 2021. 2

[35] Chunhui Zhang, Li Liu, Yawen Cui, Guanjie Huang, Weilin Lin, Yiqian Yang, and Yuehong Hu. A comprehensive survey on segment anything model for vision and beyond, 2023. 3

[36] Jiaming Zhang, Chaoxiang Ma, Kailun Yang, Alina Roitberg, Kunyu Peng, and Rainer Stiefelhagen. Transfer beyond the field of view: Dense panoramic semantic segmentation via unsupervised domain adaptation. *IEEE Transactions on Intelligent Transportation Systems*, 23(7):9478–9491, 2021. 1, 2, 6

[37] 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. 1, 2, 6, 7

[38] 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. 2

[39] Yichi Zhang and Rushi Jiao. How segment anything model (sam) boost medical image segmentation? *arXiv preprint arXiv:2305.03678*, 2023. 2

[40] Yang Zhang, Philip David, and Boqing Gong. Curriculum domain adaptation for semantic segmentation of urban scenes. In *Proceedings of the IEEE international conference on computer vision*, pages 2020–2030, 2017. 2

[41] 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, 2, 6

[42] 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. 1, 2, 6, 7

[43] Xu Zheng, Pengyuan Zhou, Athanasios Vasilakos, and Lin Wang. Semantics, distortion, and style matter: Towards source-free uda for panoramic segmentation, 2024. 2

[44] Jinjing Zhu, Haotian Bai, and Lin Wang. Patch-mix transformer for unsupervised domain adaptation: A game perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3561–3571, 2023. 2

[45] Jinjing Zhu, Yunhao Luo, Xu Zheng, Hao Wang, and Lin Wang. A good student is cooperative and reliable: Cnn-transformer collaborative learning for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11720–11730, 2023. 2