remote sensing

MDPI

Technical Note

Exploring Semantic Prompts in the Segment Anything Model for Domain Adaptation

Ziquan Wang, Yongsheng Zhang, Zhenchao Zhang *, Zhipeng Jiang, Ying Yu, Li Li and Lei Li

School of Geospatial Information, PLA Strategic Support Force Information Engineering University, Zhengzhou 450001, China; aresdrw@163.com (Z.W.); yszhang2001@vip.163.com (Y.Z.); jiangzp0803@163.com (Z.J.); yuying5559104@163.com (Y.Y.); lili315114@163.com (L.L.); 3110100798@zju.edu.cn (L.L.)

* Correspondence: zhzhc_1@163.com; Tel.: +86-150-9330-3012

Abstract: Robust segmentation in adverse weather conditions is crucial for autonomous driving. However, these scenes struggle with recognition and make annotations expensive, resulting in poor performance. As a result, the Segment Anything Model (SAM) was recently proposed to finely segment the spatial structure of scenes and to provide powerful prior spatial information, thus showing great promise in resolving these problems. However, SAM cannot be applied directly for different geographic scales and non-semantic outputs. To address these issues, we propose SAM-EDA, which integrates SAM into an unsupervised domain adaptation mean-teacher segmentation framework. In this method, we use a "teacher-assistant" model to provide semantic pseudo-labels, which will fill in the holes in the fine spatial structure given by SAM and generate pseudo-labels close to the ground truth, which then guide the student model for learning. Here, the "teacher-assistant" model helps to distill knowledge. During testing, only the student model is used, thus greatly improving efficiency. We tested SAM-EDA on mainstream segmentation benchmarks in adverse weather conditions and obtained a more-robust segmentation model.

Keywords: segment anything model (SAM); unsupervised domain adaptation; semantic road scene segmentation

Citation: Wang, Z.; Zhang, Y.; Zhang, Z.; Jiang, Z.; Yu, Y.; Li, L.; Li, L. Exploring Semantic Prompts in the Segment Anything Model for Domain Adaptation. *Remote Sens.* **2024**, *16*, 758. https://doi.org/10.3390/rs16050758

Academic Editors: Qian Du, Jiaojiao Li, Wei Li, Jocelyn Chanussot, Rui Song, Yunsong Li and Bobo Xi

Received: 20 November 2023
Revised: 1 February 2024
Accepted: 14 February 2024
Published: 21 February 2024

1. Introduction

The semantic segmentation [1–7] of road scenes is important for autonomous driving [5], particularly during scene data analyses and behavior decision-making [8]. This technology also has good applications in motion control planning [9,10] and multi-sensor fusion processing [11]. Furthermore, over the past decade, we have seen tremendous advancements in semantic segmentation technology [7,12–15]. Currently, intelligent semantic segmentation algorithms can even outperform humans in recognizing clear scenes [15]. However, these works mostly ignore the deterioration of image quality caused by adverse weather conditions such as fog, rain, and snow [16]. This leads to an obvious performance decline. Unfortunately, the reliable and safe operation of intelligent systems requires the underlying recognition processes to be highly robust under these adverse conditions. Thus, this issue is receiving increasing attention now.

Adverse weather conditions bring two main challenges to semantic segmentation. Firstly, important objects become blurred, which leads to higher uncertainty in the outputs of these intelligent algorithms. Although some studies have tried to restore these images [17] and have attempted to convert them into images with clear scenes, a domain gap still exists. Secondly, annotating these scenarios is more difficult than annotating clear ones, making it expensive to use supervised algorithms. Therefore, many studies have adopted unsupervised domain adaptation (UDA) strategies [18–20] in an attempt to transfer segmented knowledge from a clear annotated source domain to adverse weather scenes (the target domain). However, in the transfer process, a domain gap in the UDA

methods inevitably leads to information loss, resulting in imprecise segmentation in the target domain scenario.

Recently, the Segment Anything Model (SAM) [1] has attracted much attention as it uses massive amounts of data to pre-train and conduct self-supervised learning, acquiring an extremely strong generalization ability. Such a generalization ability enables SAM to be directly applied to various vision-based tasks without task-oriented training, including camouflaged object detection [21] and image in-painting [22]. Concretely, SAM can finely segment all objects in an image, thus providing powerful prior spatial structure information. Even in adverse conditions, SAM remains robust [23]. Thanks to SAM's generalization ability, SAM-DA [24] can make predictions from nighttime images and has a large number of samples for training, which greatly improves the performance of the model. Thus, we can assume that applying SAM's spatial structure information to UDA methods, i.e., adding a powerful supervision signal to the UDA framework, will be beneficial.

However, currently, SAM cannot be integrated directly into the UDA framework for three main reasons: (1) As mentioned above, SAM is not a task-oriented model, and a well-designed access plugin is needed to adapt it to semantic segmentation tasks. (2) Limited by its computing power, SAM is difficult to mount on the platform of a vehicle. (3) The operational speed of SAM is very slow and is insufficient when applied to real scenarios. For problem (1), the SSA [25] method can be used to fuse the spatial structure information generated by SAM with the semantic information generated by a segmentation model. However, the SSA method exacerbates the problem of slow operation, taking 40–60 s to complete segmentation for just one image, and its original semantic branch has not been trained to adapt to adverse weather conditions, resulting in inaccurate information and, therefore, producing unsatisfactory results. For problems (2) and (3), some scholars put forward Fast-SAM [26] and Faster-SAM [26], which have greatly improved the operational efficiency of SAM and can be deployed from mobile terminals, thus further adding significance to the research in this paper, **as is shown in Figure 1**.

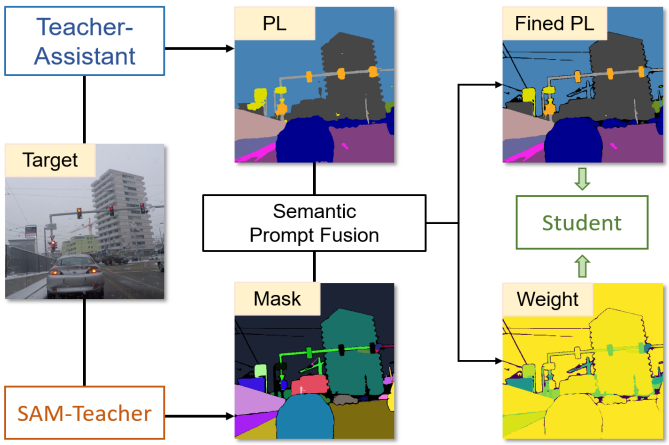

Figure 1. The main idea behind the proposed method. For images from the target domain, the teacher-assistant model and SAM-teacher generate semantic segmentation masks (called "semantic prompts") and spatial structure masks, respectively, and, then, use the algorithm mentioned in Section 2.2 for fusion. Due to SAM's strong generalization ability, this step can produce pseudo-labels that are more consistent with real scene distributions; so, the student model can completely explore the target domain knowledge, similar to the method of supervised learning.

To address the above issues, we propose a SAM-enhanced UDA method called **SAM-EDA as shown in Figure 2**, aiming to improve segmentation performance by utilizing the SAM knowledge while maintaining its original operational speed. Specifically, we plugged SAM (or its variants) into a mean-teacher's self-training domain adaptation architecture [19,27], dynamically carrying out SAM-enhanced learning on the target domain, as well as knowledge distillation.

The whole architecture and pipeline consist of three sub-modules: (1) the student model, (2) the teacher-assistant (TA) model, and (3) the SAM-teacher model. However, only the student segmentation model will be published for evaluation. In a single training iteration, the TA and SAM-teacher models generate semantic segmentation masks (called "semantic prompts") and spatial structure masks on the target domain, respectively, and, then, use the pseudo-label fusion algorithm mentioned in Section 2.2 for fusion. Due to SAM's strong generalization ability, this step can produce pseudo-labels that are more consistent with real scene distributions, so the student model can completely explore the target domain knowledge, similar to the method of supervised learning. After completing the training, neither the SAM-teacher nor TA models remain, thus maintaining the speed of the existing semantic segmentation network.

The contributions of this article can be summarized as follows:

(1) We propose a simple, but effective semantic filling and prompt method for SAM masks, which utilizes the output of existing semantic segmentation models to provide SAM with class information and explore methods to address the scale of the SAM segmentation results;
(2) To the best of our knowledge, we are the first to incorporate SAM into an unsupervised domain adaptation framework, which includes the SAM-teacher, teacher-assistant, and student models, achieving knowledge distillation in the case of completely inconsistent structures and output spaces between SAM and the main segmentation model, effectively improving its adaptability in adverse scenarios;
(3) Our method is applicable to different UDA frameworks and SAM variants, providing useful references for the application of large models in local professional fields.

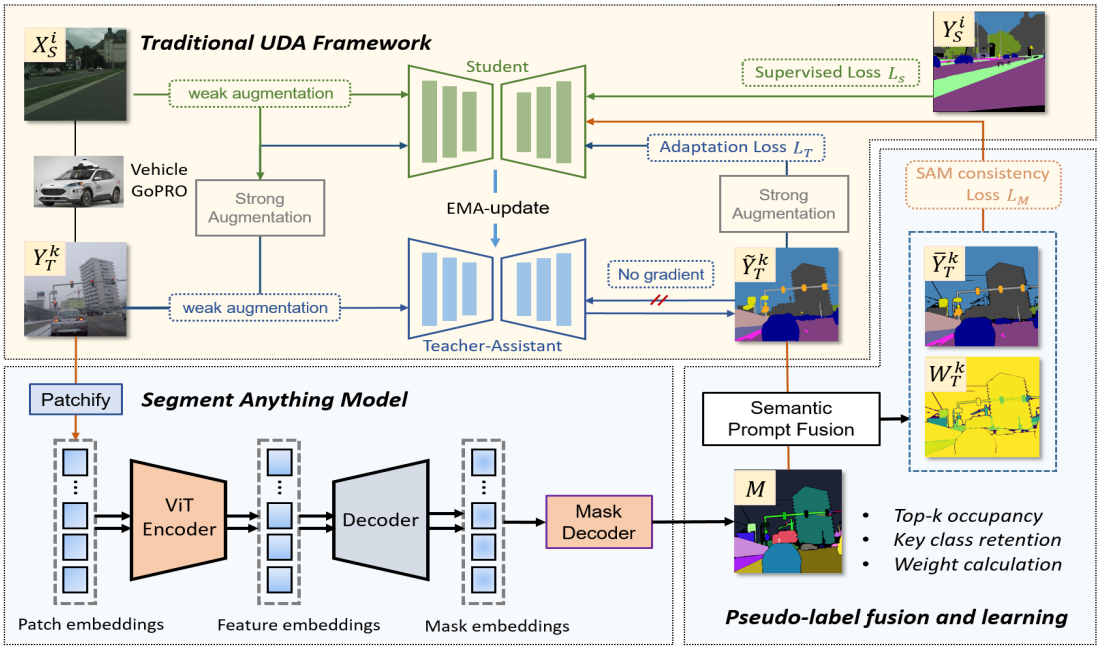

Figure 2. The pipeline of the proposed method. Both the source and target domain images used in this method were captured from a vehicle perspective camera. The target domain image Y_T was first fed into the teacher-assistant g_ϕ to generate coarse pseudo-labels $\widetilde{Y}_T$, which serve as semantic prompts. Then, Y_T was put into SAM to obtain a spatial structural segmentation map M, leveraging SAM's generalization. We merged $\widetilde{Y}_T$ and M to incorporate the semantic information. During the merging process, the top-k occupancy ratio method was mainly used to retain some key class pixels from $\widetilde{Y}_T$ while considering the holes in the SAM's missing segmentation. The weights were also calculated based on the proportion of semantic pixels to reduce the impact of uncertainty in SAM. The merged pseudo-label $\overline{Y}_T$ was close to the distribution of the real-world scene, thus enabling supervision of the student model.

2. Method

2.1. Unsupervised Domain Adaptation (UDA)

In order to perform an unsupervised domain adaptation for semantic segmentation, we utilized a student network f_θ and a teacher-assistant network g_ϕ based on the mean-teacher [19,27] pipeline. Given a set of labeled source domain data $\{(X_S^i, Y_S^i)\}_{i=1}^{N_S}$ (where Y_S^i is the pixel-wise semantic label of X_S^i), the student network directly learns from the source domain data using the cross-entropy loss function:

$$\mathcal{L}_i^{S,cls/seg} = \mathcal{H}(f_\theta(X_S^i), Y_S^i) \tag{1}$$

$$\mathcal{H}(\widetilde{y}, y) = -\sum_{h=1}^{H}\sum_{w=1}^{W}\sum_{c=1}^{C} y_{hwc} \log \widetilde{y}_{hwc} \tag{2}$$

However, models trained only on the source domain often lack generalization; thus, knowledge from the target domain needs to be extracted with N_T unlabeled images $\{(X_T^k)\}_{k=1}^{N_T}$. In our UDA pipeline, the teacher-assistant network g_ϕ needs to make predictions using the target domain images and needs to generate pseudo-labels $\{(\widetilde{Y}_T^k)\}_{k=1}^{N_T}$, so the learning loss function based on the pseudo-labels can be denoted as L_k^T, which is similar to the supervised $L_i^{S,cls/seg}$:

$$\mathcal{L}_k^T = \mathcal{H}(g_\phi(X_T^k), \widetilde{Y}_T^k) \tag{3}$$

The pseudo-labels generated by g_ϕ are often inaccurate (especially during the early stages of training), so it is necessary to set a dynamic weight λ to balance the impact of noise in the pseudo-labels. Generally, λ is set as the confidence pixel ratio exceeding a certain threshold τ:

$$\lambda_T^k = \frac{\sum_{p=1}^{H\times W}\left[\max_{c'} g_\phi(X_T^k)^{(p,c')} \geq \tau\right]}{H \times W} \tag{4}$$

Finally, the total loss function of our UDA architecture is the weighted sum of source domain loss and target domain loss:

$$\min_\theta \frac{1}{N_S}\sum_{i=1}^{N_S}\mathcal{L}_i^S + \frac{1}{N_T}\sum_{k=1}^{N_T}\lambda^T\mathcal{L}_k^T \tag{5}$$

2.2. Semantic Prompt Fusion and Learning

After standard UDA loss computation, we employed SAM [1] (or its variant) to make additional predictions on the target domain image X_T^k. Taking the standard SAM as an example, the input image was first patchified, automatically calculating the points of each patch as prompts. Then, SAM used a ViT-based [28] encoder and decoder head to obtain a feature embedding and a mask embedding. Finally, a mask decoder head identified several masks M without semantic information. Let h_φ denote SAM-series used in our pipeline. This process can be described as follows:

$$\{M_j\}_{j=1}^{N_k^m} = h_\varphi(X_T^k) \tag{6}$$

The distribution of M closely matches real-world scenarios, but it requires semantic information from pseudo-label $\widetilde{Y}_T$, referred to as a "semantic prompt". For a single mask M_j, the class ID can be obtained by calculating the most-frequent category ID within the corresponding region in $\widetilde{Y}_T$:

$$M_j^{cls} = \arg\max_c(count(M_j \odot \widetilde{Y}_T^k)_{p=c}) \tag{7}$$

However, existing dataset label systems are often restricted to the broadest instance-level labels (such as cars, buses, buildings, etc.), while SAM's segmentation has multi-scale outputs (e.g., window, car, etc.). This leads to some errors when preserving the SAM segmentation masks (see Section 4.2), meaning that certain parts of the SAM output are not fully representative of their objects, which makes it difficult to avoid using general rules. To mitigate this, we calculated the weights for each mask M_j based on the maximum occupied pixel's class ID proportion:

$$W_j^{cls} = \frac{\text{sum}(count(M_j \odot \widetilde{Y}_T^k)_{p=c})}{|M_j|} \tag{8}$$

Then, due to holes being present in small objects when using the SAM and the potential confusion between similar classes (such as roads and sidewalks), it is important to preserve some pixels to identify key classes. Taking the Cityscapes dataset [12] as an example, we selected a set of classes among $[0, 19]$, denoted as set K, through empirical judgment. The final obtained pseudo-label is a combination of the masks M_j, along with the inclusion of key class pixels from $\widetilde{Y}_T$:

$$\overline{Y}_T^k = \bigcup_{j=1}^{n_k^m} M_j^{cls} \cup \widetilde{Y}_T^k[c = c_d] \quad c_d \in K \tag{9}$$

For the mask filled using Equation (7), the weights are determined using Equation (8). For the remaining parts, the weights (which will participate in the loss function) were uniformly set to 1. The final weight matrix is as follows:

$$\overline{W}_T^k = \bigcup_{j=1}^{n_k^m} W_j^{cls} \cup \mathbf{1} \odot \widetilde{Y}_T^k[c = c_d] \quad c_d \in K \tag{10}$$

Thus, we can obtain the pseudo-labels enhanced by SAM, which can be used to construct a loss function similar to that in Equation (1), namely $\mathcal{L}_M = \mathcal{H}(f_\theta(X_T^k), \overline{Y}_T^k)$. Consequently, the final loss function is

$$\min_\theta \frac{1}{N_S} \sum_{i=1}^{N_S} \mathcal{L}_i^S + \frac{1}{N_T} \sum_{k=1}^{N_T} (\lambda^T \mathcal{L}_k^T + W_T^k \mathcal{L}_M) \tag{11}$$

3. Results

3.1. Implementation Details

3.1.1. Adverse Condition Semantic Segmentation Dataset

We used the Cityscapes dataset [12] as the source domain for training, which includes 2975 training images. The candidate target domain includes four different datasets—ACDC [16], Foggy Driving [29] and Foggy Driving Dense [30], Rainy Cityscapes [31], and Dark-Zurich (DZ) [32]—covering images with adverse conditions such as foggy, rainy, snowy, and nighttime scenes. Among them, ACDC contains 1600 images for training and 400 images for validation. Dark-Zurich contains 2416 training images and 151 test images. All datasets were labeled according to the Cityscapes standard, which includes 19 categories.

3.1.2. SAM-EDA Parameters

For the UDA architecture, we used DAFormer [19] as the baseline. Both the teacher-assistant and student models were SegFormer [7] with an MiT-B5 backbone. We followed DAFormer to set the EMA parameter $\alpha = 0.99$ and the confidence threshold $\tau = 0.968$. For the SAM mask generator, we used the largest SAM-ViT-H [1,28] and set the prediction IoU threshold δ_{iou} to 0.8. We also set the stability score threshold δ_{sta} to 0.8 and the minimum mask region $r_{\min}$ to 50 pixels. The settings of the SAM parameters directly affect the

quality and quantity of segmentation masks and determine the geographical scale. All the experiments were conducted on a Tesla v100 graphic card with 32 GB of graphic memory, equipped with CUDA 10.2 and cudnn 7.6.5.

3.2. Performance Comparison

We compared our methods with prominent UDA methods for four kinds of adverse conditions, as well as with the segmentation method SSA [25] combined with SAM application. For foggy scenes, we compared CuDA-Net [20] and FIFO [33]; for nighttime scenes, we compared VBLC [34] and GCMA [35]. These methods are all specialized for specific scenes. As for universal domain adaptation methods, we compared DAFormer [19], CumFormer [36], and the SSA method combined with SAM. For the SSA method, we provide the results using different extractors (ViT-B, ViT-L, and ViT-H). All performance comparisons are shown in Table 1 and Figure 3. We not only provide comprehensive performance comparisons for each method, but also present their runtime and memory consumption. All evaluation metrics were calculated on the validation sets of each dataset. In Table 2, we show the improvement of our method to different UDA strategies. In Table 3 and Figure 4, we show the influence of different fusion methods between SAM-generated masks and original pseudo labels. In Table 4, we show the performance of replacing the original SAM to its variants.

Table 1. Performance comparison. Experiments were conducted on the ACDC, Foggy Driving, Foggy Driving Dense, Rainy Cityscapes, and Dark-Zurich validation sets and measured with the mean intersection over union (mIoU %) over all classes.

Model	Pub/Year	Backbone	Dataset								Speed/FPS	GPU/GB
			Fog			Rain		Snow	Night			
			ACDC-f	FD	FDD	ACDC-r	Rain-CS	ACDC-s	ACDC-n	DZ		
DAFormer [19]	CVPR 2022	SegFormer [7]	63.41	47.32	39.63	48.27	75.34	49.19	**46.13**	**43.80**	6–10	Train: 16 GB Test: 8 GB
CuDA-Net [20]	CVPR 2022	DeepLabv2 [13]	68.59	53.50	48.20	48.52	69.47	47.20	-	-		
FIFO [33]	CVPR 2022	Refinelw-101 [14]	70.36	50.70	48.90	-	-	-	-	-		
CumFormer [36]	TechRXiv 2023	SegFormer	74.92	56.25	**51.91**	57.14	79.34	62.42	44.75	43.20		
VBLC [34]	AAAI 2023	SegFormer	-	-	-	-	**79.80**	-	-	44.41		
GCMA [35]	ICCV 2019	DeepLabv2	-	-	-	-	-	-	-	42.01		
SegFormer (cs)	NeurIPS 2021	-	64.74	46.06	33.15	40.62	68.31	42.03	26.61	23.43	6–10	-
SSA + SAM + SegFormer	arXiv 2023 Github 2023	ViT-B [28]	60.57	39.02	25.33	43.17	67.51	42.93	24.97	22.36	<0.1	Train: 8–48 GB Test: 16–24 GB
		ViT-L [28]	66.78	48.02	31.33	52.94	68.69	51.47	27.69	26.73		
		ViT-H [28]	68.16	50.89	33.72	54.39	70.27	53.32	29.60	28.92		
OneFormer (cs) [15]	arXiv 2022	-	72.31	51.33	44.31	56.72	74.96	55.13	32.41	26.74	4–5	-
SSA + SAM + OneFormer	arXiv2023 GitHub2023	ViT-B	69.13	46.97	41.96	58.77	73.03	57.14	36.78	28.96	<0.1	Train: 8–48 GB Test: 16–24 GB
		ViT-L	75.94	53.14	46.78	64.25	75.62	64.21	40.14	34.25		
		ViT-H	77.87	55.61	48.41	69.25	76.31	66.22	41.22	37.43		
SAM-EDA(Ours)	-	ViT-B	68.10	50.74	43.66	54.20	71.01	55.47	33.62	27.63	6.7	Train: 8–48 GB Test: 8 GB
		ViT-L	75.30	55.49	46.98	64.68	73.41	58.12	41.30	35.45		
		ViT-H	**78.25**	**56.37**	51.25	**69.38**	76.63	**68.17**	43.15	42.63		

Table 2. SAM-EDA for UDA methods. Experiments were conducted on the ACDC-Fog validation set and measured with the mean intersection over union (mIoU %) over all classes.

UDA Method	w/o SAM-EDA	w/ SAM-EDA	Diff.
DACS [37]	61.08	64.28	+3.20
ProDA [18]	65.17	68.74	+3.57
DAFormer [19]	67.93	71.61	+3.68
CuDA-Net [20]	68.56	72.37	+3.81
CumFormer [36]	74.92	77.89	+2.97

Table 3. Different semantic prompt fusion methods. Experiments were conducted on the ACDC and Dark-Zurich validation set and measured with the mean intersection over union (mIoU %) over all classes.

Method/Datasets	ACDC-F				Dark-Z	Mean	Gain (mIoU)
	Fog	Rain	Snow	Night			
DAFormer [19]	63.41	48.27	49.19	46.13	43.80	50.16	+0.00
IoU [38]	55.19	41.58	42.17	39.48	27.66	41.22	−8.94
SSA (SegFormer) [25]	68.16	54.39	53.32	29.60	28.92	46.88	−3.28
SAM-EDA w/o Weight	74.02	65.17	62.74	38.74	39.91	56.12	+5.96
SAM-EDA w/ Weight	78.25	69.38	68.17	43.15	42.63	60.32	+10.16

Table 4. SAM-EDA for SAM variants. Experiments were conducted on the ACDC-Rain validation set and measured with the mean intersection over union (mIoU %) over all classes.

	Performance	Time (s/iter)	Memory (GB)
SAM [1]	69.38	10	8–48
Fast-SAM [39]	68.22	0.5	16
Faster-SAM [26]	68.87	0.3	16

We found that, in bright scenes, such as foggy, rainy, and snowy ones, the SAM-enhanced algorithm outperformed the UDA algorithms. The SSA method performed better than DAFormer, and some methods even outperformed CumFormer, which was newly proposed by the authors, but our SAM-EDA was better than the SSA method. This is because SAM demonstrated strong generalization in bright scenes, providing sharper contour branches. Additionally, the teacher-assistant model can generate relatively accurate pseudo-labels, contributing to better fusion. For night scenes, however, the SAM itself has a significant bias (which will be shown in Section 4.2, thereby reducing the overall performance. However, our SAM-EDA still outperformed the two SSA algorithms for night scenes. Since we only kept the student model, the testing speed and memory consumption were the same as the fast SegFormer. In Figure 3, we show the qualitative comparison. Due to space limitations, we only show the results of ACDC. Based on our method, more-precise segmentation results were obtained in categories such as poles, traffic lights, and traffic signs with obvious shapes.

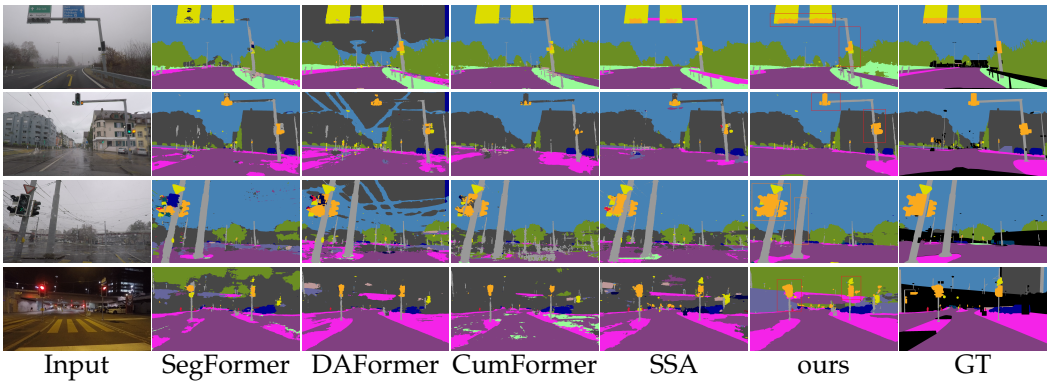

Figure 3. A qualitative comparison with other methods. From top to bottom, there are foggy, rainy, snowy, and nighttime scenes. Based on our method, more-precise segmentation results are obtained in the categories poles, traffic lights, and traffic signs with obvious shapes.

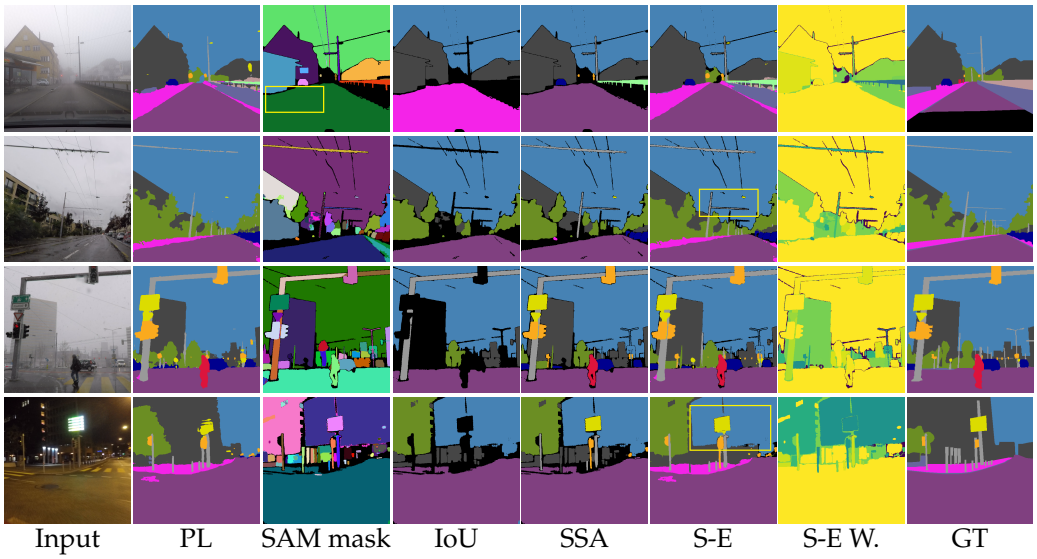

Figure 4. Different pseudo-label fusion methods. From left to right are the target domain image, the original pseudo-label (PL) generated by the teacher-assistant model, the original masks generated by the SAM, the pseudo-label fused using the IoU method, the SSA method, our SAM-EDA (S-E) method, SAM-EDA's weight, and the ground truth (GT).

4. Discussion

4.1. SAM-EDA for Different UDA Methods

We used the pseudo-labels generated by the teacher-assistant model as semantic prompts for filling in SAM's masks. In fact, SAM-EDA is suitable for any UDA segmentation method that utilizes pseudo-labels for self-training. We conducted ablation experiments on the ACDC-Fog validation set. Table 2 demonstrates the enhancement of different methods by SAM-EDA. We found that SAM-EDA can not only improve classic UDA methods (e.g., DACS [37], ProDA [18], and DAFormer [19]), but also improve methods specific to adverse scenes (CuDA-Net [20] and CumFormer [36]) by approximately 3%, indicating that SAM's information is generalizable. This shows that SAM-EDA is a good plugin, and through data-side processing, complex knowledge distillation or fine-tuning operations can be avoided, thus taking advantage of both SAM and domain-specific models.

4.2. Influence of Different Pseudo-Label Fusion Methods

Different semantic prompt fusion methods matter. We chose as many comprehensive fusion strategies as possible and present them in Figure 4. From left to right are the target domain image, the original pseudo-label generated by the teacher-assistant model, the original masks generated by SAM, the pseudo-label fused using the IoU method [38], the SSA method [25], our SAM-EDA method, and SAM-EDA's weight. From top to bottom are the four adverse-condition scenes. Among the three label fusion strategies, the simplest one is to directly assign the class that has the largest intersection over union (IoU) between the mask and category ID layer [38], which was successfully applied in weakly supervised semantic segmentation and saliency detection. However, this approach led to many holes (black areas in the fourth column of Figure 4). This is because the semantic segmentation task is at the "category level", while the SAM masks are at the instance level. When calculating the IoU, the instance-level mask takes the class-level label as the denominator, making the calculation ineffective. For example, if there are three cars in the image, in the "car" category layer, the pixels of the three cars will all be taken into account. Therefore, the proportion of pixels belonging to the "car" class in the mask of a car instance will decrease to 1/3 or 1/2 of the original proportion. If there are other classes present in the current area, it is likely that this area will be misclassified into another class. The SSA method relies entirely on the SAM mask and assigns instance-level pixel labels to all the masks output by SAM. This ensures that each mask has a definitive category label and does not generate

large areas of holes. However, if the segmentation by SAM is inaccurate, it will directly result in large areas of errors.

When dealing with the SAM masks, we identified three shortcomings. Firstly, SAM struggled to differentiate classes with similar features, such as roads and sidewalks. In all scenarios, SAM uses the same mask for roads and sidewalks. This is unacceptable for autonomous driving. Secondly, SAM has difficulty distinguishing walls from railings or simply does not recognize them as objects, which could also be fatal for autonomous driving. Lastly, SAM performed poorly in nighttime conditions. For example, in the fourth row of Figure 4, SAM mistakenly assigned large areas of buildings to the sky, leading to errors in the label fusion region and undermining the performance brought by the original pseudo-labels.

To address these issues, we retained some critical categories from the original pseudo-labels (such as sidewalks, walls, and fences) to counter SAM's shortcomings. Then, we allowed the masks to be assigned to incorrect classes (which is difficult to avoid), and we calculated the weights for each mask and reduced them in the loss function, thus effectively optimizing the SSA method. In Table 3, we show the impact of different label fusion strategies. As seen, the SAM-EDA method, which incorporates weights, achieved real improvements and outperformed the SSA method and the case without weights.

4.3. SAM-EDA for SAM Variants

SAM-EDA is also applicable to SAM variants with different numbers of parameters, with the potential to accelerate training. In Table 4, we replaced SAM with the lighter Fast-SAM [39] and Faster-SAM [26], significantly reducing the duration and memory usage of each iteration. In the standard SAM-EDA, we do not need to include SAM in the final segmentation model, so different SAM variants have little impact. However, the emergence of Faster-SAM undoubtedly provided a better option for future methods to include SAM.

4.4. Influence of SAM's Hyper-Parameters

SAM's hyper-parameters are related to the quality, density, and porosity of the generated masks. We conducted tests on the effectiveness of two hyper-parameters: the prediction IoU threshold δ_{iou} and the stability score threshold δ_{sta} (Figure 5). The higher they were set, the more precise the mask contours, but the fewer the masks. We conducted separate experiments on the ACDC-Rain validation set and found that the best results were achieved when $\delta_{iou} = \delta_{sta} = 0.8$. This indicates that we need a stable quantity of masks to cover the entire image during label fusion rather than solely focusing on the quality of the masks.

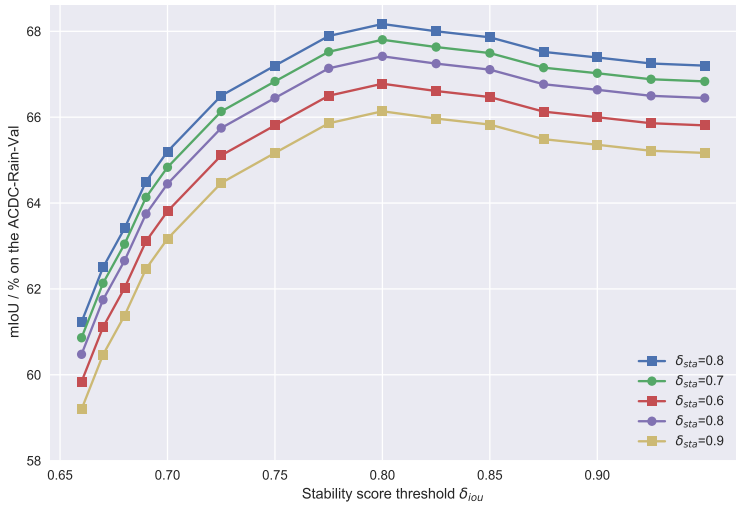

Figure 5. Influence of SAM's hyper-parameters. High δ_{iou} and δ_{sta} both result in performance degradation, and we found that the best results were achieved at $\delta_{iou} = \delta_{sta} = 0.8$.

5. Conclusions

We have presented SAM-EDA, a universal framework for using SAM in unsupervised semantic segmentation tasks. This method utilizes pseudo-labels generated by specific semantic segmentation models as prompts to fill in the spatial structure of SAM segmentation, thereby obtaining a more-accurate probability distribution of scene segmentation. The most-significant contribution of our method is the introduction of a more-accurate and fault-tolerant semantic prompt fusion approach. It can integrate the spatial structure provided by SAM with the semantic discernment generated by the original segmentation network. Our experiments showed that our method achieved better performance on semantic segmentation benchmarks under several adverse imaging conditions. Moreover, it can be implemented in a plug-and-play manner to enhance any unsupervised semantic segmentation algorithm based on pseudo-labels. After introducing a lightweight variant of SAM, our method obtained the ability to perform near real-time training and testing. We also explored the hyper-parameters of SAM.

The universality and generalizability of SAM are valuable resources. In future research, we plan to introduce SAM into tasks such as Test Time Adaptation, serving as a spatial structure anchor to combat the catastrophic forgetting that may occur during prolonged adaptation processes of the model.

Author Contributions: Conceptualization, Z.W. and Z.Z.; methodology, Z.W., Z.Z. and Z.J.; software, Z.W. and Z.J.; validation, Z.W., Y.Z. and Y.Y.; formal analysis, Y.Z. and L.L. (Li Li); investigation, Z.W., Z.Z., Y.Y. and L.L. (Li Li); data curation, Z.Z., Z.J., Y.Y., L.L. (Li Li) and L.L. (Lei Li); writing—original draft preparation, Z.W., Z.J. and L.L. (Lei Li); writing—review and editing, Z.W., Z.Z., Z.J., L.L. (Li Li) and L.L. (Lei Li); visualization, Y.Z. and Z.Z.; supervision, Y.Z. and Z.Z.; project administration, Y.Z.; funding acquisition, Y.Z. and Y.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China under Grant 42071340 and a program of the Song Shan Laboratory (managed by the Major Science and Technology Department of Henan Province) under Grant 2211000211000-01 and 2211000211000-04.

Data Availability Statement: The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author/s.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A.C.; Lo, W.Y.; et al. Segment Anything. *arXiv* **2023**, arXiv:2304.02643.
2. Šarić, J.; Oršić, M.; Šegvić, S. Panoptic SwiftNet: Pyramidal Fusion for Real-Time Panoptic Segmentation. *Remote Sens.* **2023**, *15*, 1968. [CrossRef]
3. Lv, K.; Zhang, Y.; Yu, Y.; Zhang, Z.; Li, L. Visual Localization and Target Perception Based on Panoptic Segmentation. *Remote Sens.* **2022**, *14*, 3983. [CrossRef]
4. Dai, Y.; Li, C.; Su, X.; Liu, H.; Li, J. Multi-Scale Depthwise Separable Convolution for Semantic Segmentation in Street–Road Scenes. *Remote Sens.* **2023**, *15*, 2649. [CrossRef]
5. Liu, Q.; Dong, Y.; Jiang, Z.; Pei, Y.; Zheng, B.; Zheng, L.; Fu, Z. Multi-Pooling Context Network for Image Semantic Segmentation. *Remote Sens.* **2023**, *15*, 2800. [CrossRef]
6. Sun, Q.; Chao, J.; Lin, W.; Xu, Z.; Chen, W.; He, N. Learn to Few-Shot Segment Remote Sensing Images from Irrelevant Data. *Remote Sens.* **2023**, *15*, 4937. [CrossRef]
7. Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J.M.; Luo, P. SegFormer: Simple and efficient design for semantic segmentation with transformers. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 12077–12090.
8. Yang, W.; Wang, S.J.; Khanna, P.; Li, X. Pattern Recognition Techniques for Non Verbal Human Behavior (NVHB). *Pattern Recognit. Lett.* **2019**, *125*, 684–686. [CrossRef]
9. Chen, G.; Hua, M.; Liu, W.; Wang, J.; Song, S.; Liu, C.; Yang, L.; Liao, S.; Xia, X. Planning and tracking control of full drive-by-wire electric vehicles in unstructured scenario. *Proc. Inst. Mech. Eng. Part D J. Automob. Eng.* **2023**, 09544070231195233. [CrossRef]
10. Liu, W.; Hua, M.; Deng, Z.; Meng, Z.; Huang, Y.; Hu, C.; Song, S.; Gao, L.; Liu, C.; Shuai, B.; et al. A Systematic Survey of Control Techniques and Applications in Connected and Automated Vehicles. *IEEE Internet Things J.* **2023**, *10*, 21892–21916. [CrossRef]

11. Meng, Z.; Xia, X.; Xu, R.; Liu, W.; Ma, J. HYDRO-3D: Hybrid Object Detection and Tracking for Cooperative Perception Using 3D LiDAR. *IEEE Trans. Intell. Veh.* **2023**, *8*, 4069–4080. [CrossRef]
12. 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 Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 3213–3223.
13. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 834–848. [CrossRef]
14. Lin, G.; Milan, A.; Shen, C.; Reid, I. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1925–1934.
15. Jain, J.; Li, J.; Chiu, M.; Hassani, A.; Orlov, N.; Shi, H. OneFormer: One Transformer to Rule Universal Image Segmentation. *arXiv* **2022**, arXiv:2211.06220.
16. Sakaridis, C.; Dai, D.; Van Gool, L. ACDC: The adverse conditions dataset with correspondences for semantic driving scene understanding. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 10765–10775.
17. Ren, W.; Ma, L.; Zhang, J.; Pan, J.; Cao, X.; Liu, W.; Yang, M.H. Gated fusion network for single image dehazing. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 3253–3261.
18. Zhang, P.; Zhang, B.; Zhang, T.; Chen, D.; Wang, Y.; Wen, F. Prototypical pseudo label denoising and target structure learning for domain adaptive semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 12414–12424.
19. Hoyer, L.; Dai, D.; Van Gool, L. Daformer: Improving network architectures and training strategies for domain-adaptive semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 9924–9935.
20. Ma, X.; Wang, Z.; Zhan, Y.; Zheng, Y.; Wang, Z.; Dai, D.; Lin, C.W. Both style and fog matter: Cumulative domain adaptation for semantic foggy scene understanding. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 18922–18931.
21. Tang, L.; Xiao, H.; Li, B. Can SAM Segment Anything? When SAM Meets Camouflaged Object Detection. *arXiv* **2023**, arXiv:2304.04709.
22. Wang, X.; Wang, W.; Cao, Y.; Shen, C.; Huang, T. Images speak in images: A generalist painter for in-context visual learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 6830–6839.
23. Shan, X.; Zhang, C. Robustness of Segment Anything Model (SAM) for Autonomous Driving in Adverse Weather Conditions. *arXiv* **2023**, arXiv:2306.13290.
24. Yao, L.; Zuo, H.; Zheng, G.; Fu, C.; Pan, J. SAM-DA: UAV Tracks Anything at Night with SAM-Powered Domain Adaptation. *arXiv* **2023**, arXiv:2307.01024.
25. Chen, J.; Yang, Z.; Zhang, L. Semantic Segment Anything. 2023. Available online: https://github.com/fudan-zvg/Semantic-Segment-Anything (accessed on 5 May 2023).
26. Zhang, C.; Han, D.; Qiao, Y.; Kim, J.U.; Bae, S.H.; Lee, S.; Hong, C.S. Faster Segment Anything: Towards Lightweight SAM for Mobile Applications. *arXiv* **2023**, arXiv:2306.14289.
27. Tarvainen, A.; Valpola, H. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Volume 30.
28. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16 $\times$ 16 Words: Transformers for Image Recognition at Scale. *arXiv* **2020**, arXiv:2010.11929.
29. Sakaridis, C.; Dai, D.; Van Gool, L. Semantic foggy scene understanding with synthetic data. *Int. J. Comput. Vis.* **2018**, *126*, 973–992. [CrossRef]
30. Sakaridis, C.; Dai, D.; Hecker, S.; Van Gool, L. Model adaptation with synthetic and real data for semantic dense foggy scene understanding. In Proceedings of the of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 687–704.
31. Lin, H.; Li, Y.; Fu, X.; Ding, X.; Huang, Y.; Paisley, J. Rain o'er me: Synthesizing real rain to derain with data distillation. *IEEE Trans. Image Process.* **2020**, *29*, 7668–7680. [CrossRef]
32. Dai, D.; Gool, L.V. Dark Model Adaptation: Semantic Image Segmentation from Daytime to Nighttime. *arXiv* **2018**, arXiv:1810.02575.
33. Lee, S.; Son, T.; Kwak, S. Fifo: Learning fog-invariant features for foggy scene segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 18911–18921.
34. Li, M.; Xie, B.; Li, S.; Liu, C.H.; Cheng, X. VBLC: Visibility Boosting and Logit-Constraint Learning for Domain Adaptive Semantic Segmentation under Adverse Conditions. *arXiv* **2022**, arXiv:2211.12256.
35. Sakaridis, C.; Dai, D.; Gool, L. Guided Curriculum Model Adaptation and Uncertainty-Aware Evaluation for Semantic Nighttime Image Segmentation. *arXiv* **2019**, arXiv:1901.05946.

36. Wang, Z.; Zhang, Y.; Ma, X.; Yu, Y.; Zhang, Z.; Jiang, Z.; Cheng, B. Semantic Segmentation of Foggy Scenes Based on Progressive Domain Gap Decoupling. *TechRxiv* **2023**. [CrossRef]
37. Tranheden, W.; Olsson, V.; Pinto, J.; Svensson, L. Dacs: Domain adaptation via cross-domain mixed sampling. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January 2021; pp. 1379–1389.
38. Chen, T.; Mai, Z.; Li, R.; Chao, W.l. Segment anything model (sam) enhanced pseudo labels for weakly supervised semantic segmentation. *arXiv* **2023**, arXiv:2305.05803.
39. Zhao, X.; Ding, W.; An, Y.; Du, Y.; Yu, T.; Li, M.; Tang, M.; Wang, J. Fast Segment Anything. *arXiv* **2023**, arXiv:2306.12156.