Prompt What You Need: Enhancing Segmentation in Rainy Scenes with Anchor-based Prompting

Xiaoyu Guo, Xiang Wei, Qi Su, Huiqin Zhao and Shunli Zhang School of Software Engineering, Beijing Jiaotong University

Abstract—Semantic segmentation in rainy scenes is a challenging task due to the complex environment, class distribution imbalance, and limited annotated data. To address these challenges, we propose a novel framework that utilizes semi-supervised learning and pre-trained segmentation foundation model to achieve superior performance. Specifically, our framework leverages the semi-supervised model as the basis for generating raw semantic segmentation results, while also serving as a guiding force to prompt pre-trained foundation model to compensate for knowledge gaps with entropy-based anchors. In addition, to minimize the impact of irrelevant segmentation masks generated by the pre-trained foundation model, we also propose a mask filtering and fusion mechanism that optimizes raw semantic segmentation results based on the principle of minimum risk. The proposed framework achieves superior segmentation performance on the Rainy WCity dataset and is awarded the first prize in the subtrack of STRAIN in ICME 2023 Grand Challenges.

Index Terms—semantic segmentation, real-world rainy scenes, semi-supervised learning, foundation model

I. INTRODUCTION

Semantic segmentation is a critical task in computer vision that involves assigning a class label to each pixel in an image. Semantic segmentation has numerous applications, including autonomous driving [1], surveillance [2], and robotics [3]. Despite significant advancements in semantic segmentation, accurately segmenting images in complex scenarios, such as rainy scenes [4], [5], remains a formidable challenge. In general, rainy scenes introduce significant complexities due to factors such as environmental variability, class distribution imbalance, and the scarcity of annotated data. These challenges frequently result in a decline in segmentation performance, underscoring the need for enhanced methods capable of effectively addressing such specific scenarios.

Recent developments in foundation segmentation models, such as Segment Anything Model (SAM) [6] and SegGPT [7], have demonstrated impressive results on a wide range of segmentation tasks in zero-shot scenarios. However, these models do not perform up to expectations when it comes to semantic segmentation in rainy scenes (as shown in Fig. 1). These limitations prompt the necessity for a novel framework that can harness the power of pre-trained foundation models without retraining while addressing the unique challenges posed by rainy scenes.

In this paper, we devise an innovative framework that combines semi-supervised techniques with pre-trained foun-

This research was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 61906014 and 61976017.

Xiang Wei is the corresponding author.





(a) Input Image

(b) Segment Anything Model

Fig. 1. Demonstration of SAM. In sub-figure (b), the performance of SAM is influenced by environmental factors, particularly reflections in rainy scenes. Additionally, the segmentation results generated by SAM may not fully adapt to specific tasks and may also potentially ignore small entities.

dation models to effectively tackle semantic segmentation in rainy scenes with limited labeled training data. In brief, our framework consists of the following three steps:

- We first leverage semi-supervised base model U²PL [8] to provide guidance information, as it is capable of utilizing unreliable pixels for representation learning by contrastive learning. This is suitable for handling uncertainty caused by environment interference.
- We further propose to use the high-entropy regions calculated from U²PL's predictions to generate anchors for prompting SAM. This strategy enables the identification of entities heavily impacted by rainy scenes, which tend to be more challenging to classify accurately.
- Finally, we put forward a filtering and fusion mechanism that carefully utilizes the segmentation masks generated by SAM to refine the predictions made by U²PL.

The proposed framework achieves superior segmentation performance on the Rainy WCity dataset and is awarded the first prize in the Seeing Through the Rain (STRAIN) – Track 1: Semantic Segmentation under Real Rain Scene, which is part of Grand Challenges in the International Conference of Multimedia and Expo 2023. Furthermore, our framework also provides insights and inspiration for active prompting in promptable foundation models.

II. RELATED WORK

Real-world semantic segmentation in rainy scenes presents several challenges. To solve the challenges, we provide a brief overview of semi-supervised learning in semantic segmentation and foundation models for computer vision.

A. Semi-Supervised Semantic Segmentation

Semi-supervised learning [9], [10] has been widely adopted to overcome the challenge of limited labeled data. Early methods such as those proposed in [11] leverage generative adversarial networks to train on unlabeled data using an adversarial loss, thereby reducing the gap between predictions on labeled and unlabeled data. In recent years, consistency regularization [12], self-training [13], [14], and their combinations [9], [15]–[17] have become the mainstream in semi-supervised semantic segmentation. These methods aim to use unlabeled data to improve model performance while reducing the impact of label noise, such as weak and strong augmentations of the same sample. Improving the quality of trusted pseudo-labels is crucial for self-training and enhancing model performance [18]. Additionally, contrastive learning has shown promising results in semi-supervised feature extraction [19], [20]. U²PL [8] uses the value of entropy for filtering reliable pixel-wise pseudo-labels and pushes the remaining unreliable pixels to a category-wise memory bank for contrastive learning, resulting in improved segmentation performance.

Unlike existing semi-supervised learning approaches, our framework exclusively relies on the semi-supervised model as the base model. The base model guides a pre-trained segmentation foundation model to bridge the knowledge gaps between the two models, leading to improved accuracy of the segmentation outcomes.

B. Foundation Model for Computer Vision

The field of natural language processing (NLP) is being revolutionized by large language models pre-trained on web-scale datasets, e.g., ChatGPT. These models, commonly referred to as "foundation models" [21], are capable of strong zeroshot and few-shot generalization, extending their capabilities to tasks and data distributions beyond those seen during training. Similarly, beyond NLP, foundation models in the field of computer vision are also becoming increasingly popular. CLIP [22] and ALIGN [23] are examples of foundation models that adopted contrastive learning to train text and image encoders, enabling zero-shot generalization to novel visual concepts and data distributions using text prompts. While much progress has been made in vision and language encoders, many computer vision problems lack abundant training data. Recently, SAM [6] and SegGPT [7] are proposed for image segmentation, both SAM and SegGPT are promptable model and have been pre-trained on a broad dataset using a task that enables powerful generalization. Nevertheless, both SAM and SegGPT require manual examples to achieve the expected results and struggle to maintain good performance in semantic segmentation for rainy scenes, as illustrated in Fig. 1b.

Different from existing prompt methods for pre-trained segmentation foundation models, we leverage the uncertainty regions in the predictions of the semi-supervised base model to generate anchors. These anchors accurately identify the weaknesses of the semi-supervised base model and enable more effective utilization of the pre-trained foundation model's knowledge. Specifically, we use SAM in our framework.

III. METHODOLOGY

In this section, we detail our three-stage framework for semantic segmentation in rainy scenes.

A. Overall Architecture

As depicted in Figure 2, our framework comprises three main steps. First, to overcome the challenge of limited annotated data, we employ semi-supervised learning to train the base model for semantic segmentation. Next, given the difficulty in establishing accurate semantics in areas affected by rainy environments in the existing dataset, we identify the image regions affected by interference by computing the entropy values of the semi-supervised base model's predictions. We then generate anchors in these image regions and use them to prompt the SAM to make predictions, resulting in a set of predicted segmentation masks. Finally, we meticulously leverage the segmentation masks generated by SAM to refine the original predictions.

B. Semi-supervised Base Model Learning

Since only one out of five of the training data has annotated labels, we suggest using a semi-supervised semantic segmentation model for initial training. For semi-supervised semantic segmentation, generating pixel-wise pseudo-labels in rainy scenes can be highly uncertain due to the environmental interference. In the semi-supervised learning step, we follow the U²PL [8] and make suitable settings to it to adapt to semantic segmentation in rainy scenes. For annotated data, we adapt OHEM [24], which is responsible for mining difficult pixels and forces the model to focus on image regions that are affected by environmental interference, and this manner is effective for semantic segmentation in rainy scenes.

C. Anchor-based Prompting

Considering that SAM can generate more accurate semantic segmentation results leveraging coordinate points as guidance. Furthermore, from our attempts, we have discovered that identifying the coordinate points according to the weakness of semi-supervised base model is crucial while using SAM to compensate for knowledge gaps. To pinpoint the model's weakness, we utilize the entropy values of predictions, which can be formalized by Equation (1):

$$Ent_{ij} = -\sum_{k=1}^{N} p_{ijk} \log p_{ijk} \tag{1}$$

where N denotes the number of classes, and p_{ijk} stands for the k-th softmax score of the pixel in the *i*-th row and *j*-th column. By examining the entropy map displayed in Fig. 3b, we can discern that the boundaries of segmentation possess larger entropy values. The high entropy distribution caused by segmentation boundaries makes it difficult for us to pinpoint the image regions where the model is truly confused.

To reduce the interference caused by segmentation boundaries, we design a region filter, which can also be interpreted as a $w \times w$ 2-D kernel with a fixed value 1 of weights



Fig. 2. The overall process of our framework. The framework consists of three main steps. In step 1, we leverage semi-supervised learning to train a base model. In step 2, based on the predictions of the semi-supervised model, we calculate the entropy values of the predictions and remove the impact of segmentation boundaries to generate high-entropy regions, which are then treated as anchors. SAM generates semantic segmentation masks according to the generated anchors. In step 3, we optimize the predictions of the base model using the masks generated by SAM.

then attached with a binarized activator. Besides, w is an odd number to guarantee a explicit center. Equation (2) details the operations of the region filter:

$$Reg_{ij} = f(Ent_{ij}) = \mathbb{1}_{\tau} (\frac{1}{w^2} \sum_{m=-w}^{w} \sum_{n=-w}^{w} Ent_{i+m,j+n}).$$
(2)

Here, the function $f(\cdot)$ serves as a region filter, which converts the entropy map into a 0-1 mask with 1 indicating the highentropy regions. $Reg_{ij} = f(Ent_{ij})$ is the entropy value of the *i*-th row and *j*-th column pixel after being filtered and binarized. The binarized activator function with a threshold of τ is denoted by $\mathbb{1}_{\tau}(\cdot)$. If the input is greater than or equal to τ , $\mathbb{1}_{\tau}(\cdot)$ returns 1; otherwise, it returns 0. The term $1/w^2$ is used for normalization.

As illustrated in Fig. 3b and 3c, the filtered entropy map includes almost no finer segmentation boundaries, which only caused by intermediate pixels between different categories. Specifically, Fig. 3b displays the entropy map derived from the semi-supervised base model. As shown in Fig. 3c, undergoing the region filter, entropy values previously displayed by finer segmentation boundaries are no longer presented. On the contrary, regional high entropy values are retained with difficulty in classification caused by environmental interference and imbalanced class distribution.

Based on the acquired high-entropy regions, we randomly sample coordinate points as anchors for prompting SAM to compensate for the lack of knowledge in the base model. The anchors generated by our method revealed in Fig. 3d. Then, we obtain corresponding segmentation results from SAM, which are binary masks for segmented entities.

D. Segmentation Refinement

Once the segmentation masks have been obtained from the SAM, they are supplemented to refine the predictions made



(c) High-entropy Region

(d) Anchors

Fig. 3. The procedure of entropy-based anchor generation. Sub-figure (a) displays the original input image, while sub-figure (b) shows the entropy map produced by the semi-supervised base model. Sub-figure (c) highlights the regions of the image with high uncertainty after filtering using region filter and sub-figure (d) displays the sampled anchors marked with stars.

by the semi-supervised base model. However, it is important to note that not all segmentation masks generated by the SAM from the anchors are reliable, so the key issue is how to extract useful knowledge from numerous masks to refine the predictions of the semi-supervised base model. Namely, we want to minimize the risk of reducing the accuracy of the existing segmentation results made by the semi-supervised base model when using the supplementary masks.

The principle of minimum risk: Go back to Fig. 1b, the segmentation masks generated by SAM contain a lot of interference information, such as incorrect entities and errors caused by environmental factors. To improve the quality of the masks during fusion, we propose to use two hyperparameters to filter the masks. One is the softmax score α , which filters low confidence segmentation masks; the other is a considerably small area β , which ensures the entity segmented by SAM is part of the target category, e.g., in Fig. 1b, the "umbrella", "pants", and "shirt" are entities that segmented by SAM, but they belong to the "person" category. Furthermore, in case of a conflict, the priority of smaller segmented entities should be higher than that of larger entities. For instance, SAM may segment a "person" and a "bicycle" as a single entity based on some anchors, and we can correct this error by using the "person" entities segmented based on other anchors.

Algorithm 1 Segmentation Enhancement

Require: Pi	rediction y and	softmax	score p,	kernel	size	<i>w</i> ,
threshol	d of binarizatio	n $ au$, thresh	hold of so	oftmax	score	α,
and thre	shold of area β	}				
Ensure: En	hanced predicti	on y				

```
1: function SEGENHANCE(y, p, w, \tau, \alpha, \beta)
                                                                ⊳ Eq.1
2:
        Ent \leftarrow \text{computeEntropy}(p)
        Reg \leftarrow getHighEntropyRegions(Ent, \tau, w)
                                                                \triangleright Eq.2
3:
        Anc \leftarrow generateAnchors(Reg)
4:
        for Anc_i in Anc do
 5:
             M_i \leftarrow \text{segmentBySAM}(Anc_i)
 6:
 7:
        end for
        M \leftarrow \text{filterByScoreAndArea}(M, \alpha, \beta)
8:
9:
        for M_i in M do
            cls \leftarrow getModeOfIntersaction(M_i, y)
10:
             Assign cls to M_i
11:
12:
        end for
        M \leftarrow \text{sortByAreaFromHighToLow}(M)
13.
14:
        for M_i in M do
            Overwrite y according to M_i with class cls
15:
16:
        end for
        return Enhanced prediction y
17:
18: end function
```

Algorithm 1 outlines the overall process of our framework. The algorithm mainly takes two inputs, namely the prediction y and the corresponding softmax score p from the semisupervised base model. Additionally, four hyper-parameters are provided: w represents the kernel size, and τ represents the binarization threshold in Equation (2). Furthermore, α and β denote the thresholds for the softmax score and area, respectively, which are utilized to ensure the quality of the segmentation masks produced by the SAM.

To be specific, in Algorithm 1, we first calculate the entropy values based on the softmax score p by "computeEntropy" according to Equation (1). Next, we call "getHighEntropy-Regions" to apply a region filter based on Equation (2), using hyperparameters w and τ , to eliminate the influence of segmentation boundaries and retain high-entropy regions. After obtaining the high-entropy regions, we adopt "generateAnchors" to perform sampling and generate anchor points. With the help of anchors to prompt SAM, we then obtain the corresponding segmentation masks by "segmentBySAM". Once we obtain the segmentation masks M generated by SAM, we retain the low risk masks M by their softmax score and area using the function "filterByScoreAndArea", specifically, we aim to obtain entities that are segmented by SAM and have small area and high confidence. Then, for each mask M_i in M, we find the mode of intersection between M_i and y by the function "getModeOfIntersection", which serves as the class for refining y according to M_i . Next, we sort the remaining masks M by their area from high to low using the function "sortByAreaFromHighToLow", this ensures that the results of smaller entities are not covered. At last, we let M_i to determine the region that needs to be overwritten in prediction y, and write the corresponding class cls to that region.

IV. EXPERIMENTS

In this section, we report the experimental results of the proposed framework for semantic segmentation in rainy scenes on the Rainy WCity dataset.

A. Experimental Setup

Datasets: We conduct our experiments on the Rainy WCity dataset, which consists of 500 images for training and 100 images for evaluation, all with a resolution of 1920×1080 . Out of the 500 images, only 100 have pixel-wise annotations, while the remaining images are unannotated. Specifically, there are 240 raindrop images, 40 of which are annotated, 130 reflection images with 30 annotated, and 130 wiper images with 30 annotated. The dataset includes pixel-level labels for a total of 18 classes, including the background.

Implementation details: For semi-supervised base model (U²PL), the experiment runs for 200 epochs, and we choose the checkpoint of the last epoch for evaluation. The network is based on the ResNet101 and Deeplabv3+ for encoder and decoder, respectively. We dynamically drop 20% to 0% of high-entropy pixels due to the unreliability while self-training. During training, we set the batch size to 2 for each of GPU. For anchor-based prompting, we set w to 5 and τ to 1.0. We sample 1,000 anchors for each of prediction from the semi-supervised base model. For segmentation refinement, we set the threshold of softmax score α to 0.7 and threshold of area β to 20,000, respectively. The experiments were conducted using Pytorch 1.12.0 and the entire training process was completed on 8 NVIDIA 3090 GPUs.

B. Results

Table I presents the evaluation results of the top 5 teams on the Grand Challenge Proposal of the IEEE International Conference on Multimedia and Expo 2023: Seeing Through the Rain (STRAIN) - Track 1 - Semantic Segmentation under Real Rain Scene. The table displays the performance of each category, measured using the Intersection over Union (IoU) metric, as well as the overall performance for all categories, measured using the mean Intersection over Union (mIoU) metric. The results indicate that our framework has achieved state-of-the-art (SOTA) results. The ground truth of test dataset only released after finishing evaluation. Notably, the IoU value of our framework for the "person" class far exceed those of the other methods. This is mainly because SAM provide informative segmentation masks. As shown in Fig. 4b, SAM

 TABLE I

 The top 5 teams' evaluation results for the Seeing Through the Rain: Track 1 - Semantic Segmentation under Real Rain Scene, which is part of the Grand Challenge Proposals of the IEEE International Conference on Multimedia and Expo 2023

Method	road	sidew.	build	wall	fence	pole	light	sign	veget.	sky	person	rider	car	truck	bus	motorc.	bicycle	mIoU
Rank 5	87.08	38.99	65.55	69.92	28.08	16.52	6.30	35.55	73.83	85.83	28.99	3.04	64.24	4.57	18.07	6.60	26.44	38.80
Rank 4	86.29	14.46	72.63	69.72	69.36	19.82	14.25	43.89	65.62	95.14	37.28	8.34	69.17	2.66	20.56	6.48	33.81	42.91
Rank 3	84.94	24.04	75.90	47.01	68.22	27.66	40.50	51.25	83.82	94.67	41.76	0.17	78.89	0.00	70.27	27.92	36.06	50.18
Rank 2	94.62	<u>61.13</u>	84.76	84.19	78.94	39.24	58.68	<u>80.36</u>	86.46	96.36	24.95	18.49	85.26	<u>9.10</u>	70.33	<u>29.86</u>	48.78	61.85
Rank 1 (Ours)	<u>94.64</u>	53.88	<u>85.94</u>	<u>87.05</u>	78.52	<u>46.86</u>	<u>62.96</u>	79.58	<u>88.75</u>	<u>96.51</u>	<u>54.83</u>	6.63	<u>87.14</u>	7.77	<u>83.64</u>	26.72	<u>50.69</u>	64.24





(b) Filtered by Softmax Score and Area

Fig. 4. Segmentation by anchor-based prompting. The yellow region represents segmented entities by SAM. Sub-figure (a) shows the entities segmented without filtering, while sub-figure (b) shows the entities filtered based on both softmax score and area. The filtering function "filterByScoreAndArea" can be find on line 8 of Algorithm 1.

has strong segmentation capability for "person", and this advantage has been passed on to our framework. In addition, Fig. 5 visualizes part of segmentation results on test dataset.

C. Ablation Study

Table II presents the results of our ablation study, which was conducted using the ground truth provided by the official test set. Specifically, the use of SAM for enhancing the segmentation results without filtering and sorting is indicated by "ENHANCE". Additionally, the use of filtering and sorting to process segmented entities is denoted by "w/ filter" and "w/ sort", respectively, which can be found on line 8 and 13 in Algorithm 1. The results in Table II demonstrate the effectiveness of our proposed components. Moreover, we also visualize the segmentation masks generated by SAM with anchor-based prompting in Fig. 4 and make comparisons. Fig. 4b shows that anchor-based prompting can capture small and difficult-to-segment entities.

V. CONCLUSION

In conclusion, this paper presents an innovative framework for addressing the challenges of semantic segmentation in rainy scenes. Our approach successfully harnesses the power of pre-trained foundation models without retraining by utilizing the entropy-based anchors generated by a semi-supervised

 TABLE II

 The experimental results of ablation study

Method	ENHANCE	w/ filter	w/ sort	mIoU
				63.11
01100	\checkmark			62.72
Ours	\checkmark	\checkmark		62.83
	\checkmark	\checkmark	\checkmark	64.24

base model. Our experiments, conducted on the Rainy WCity dataset, demonstrate that the proposed framework effectively leverages the pre-trained segmentation foundation model, leading to superior segmentation accuracy. Furthermore, our framework also provides insights and inspiration for active prompting in promptable foundation models.

However, there are still limitations to our framework. The accuracy of refinement essentially depends on the correctness of the guidance information provided by the semi-supervised base model, and our method relies on an additional model during inference, which limits its applicability to certain scenarios. In future work, we plan to investigate how to transfer knowledge from pre-trained segmentation foundation models to models in specific domains.

REFERENCES

- [1] D. Feng, C. Haase-Schütz, L. Rosenbaum, H. Hertlein, C. Glaeser, F. Timm, W. Wiesbeck, and K. Dietmayer, "Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 3, pp. 1341–1360, 2020.
- [2] S. Piérard, A. Cioppa, A. Halin, R. Vandeghen, M. Zanella, B. Macq, S. Mahmoudi, and M. Van Droogenbroeck, "Mixture domain adaptation to improve semantic segmentation in real-world surveillance," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 22–31.
- [3] J. Li, Y. Dai, J. Wang, X. Su, and R. Ma, "Towards broad learning networks on unmanned mobile robot for semantic segmentation," in 2022 International Conference on Robotics and Automation (ICRA). IEEE, 2022, pp. 9228–9234.
- [4] Y. Zheng, X. Yu, M. Liu, and S. Zhang, "Single-image deraining via recurrent residual multiscale networks," *IEEE transactions on neural networks and learning systems*, vol. 33, no. 3, pp. 1310–1323, 2020.
- [5] X. Zhong, S. Tu, X. Ma, K. Jiang, W. Huang, and Z. Wang, "Rainy wcity: A real rainfall dataset with diverse conditions for semantic driving scene understanding," in *Proc. IJCAI Int. Joint Conf. Artif. Intell*, 2022.
- [6] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, "Segment anything," *arXiv:2304.02643*, 2023.



Fig. 5. Visualization of segmentation results. Sub-figure (b) indicates the results obtained by a pre-trained model on cityscapes, (c) shows the results of a supervised model based on OHEM, (d) shows the results of the semi-supervised model in our framework, (e) displays the final segmentation results achieved by our framework, and (f) shows the ground truth. All of methods are based on U^2PL for conducting fair comparisons.

- [7] X. Wang, X. Zhang, Y. Cao, W. Wang, C. Shen, and T. Huang, "Seggpt: Segmenting everything in context," *arXiv preprint arXiv:2304.03284*, 2023.
- [8] Y. Wang, H. Wang, Y. Shen, J. Fei, W. Li, G. Jin, L. Wu, R. Zhao, and X. Le, "Semi-supervised semantic segmentation using unreliable pseudolabels," in *Proceedings of the IEEE/CVF Conference on Computer Vision* and Pattern Recognition, 2022, pp. 4248–4257.
- [9] K. Sohn, D. Berthelot, N. Carlini, Z. Zhang, H. Zhang, C. A. Raffel, E. D. Cubuk, A. Kurakin, and C.-L. Li, "Fixmatch: Simplifying semisupervised learning with consistency and confidence," *Advances in neural information processing systems*, vol. 33, pp. 596–608, 2020.
- [10] X. Kong, X. Wei, X. Liu, J. Wang, S. Lu, W. Xing, and W. Lu, "3lpr: A three-stage label propagation and reassignment framework for classimbalanced semi-supervised learning," *Knowledge-Based Systems*, vol. 253, p. 109561, 2022.
- [11] S. Mittal, M. Tatarchenko, and T. Brox, "Semi-supervised semantic segmentation with high-and low-level consistency," *IEEE transactions* on pattern analysis and machine intelligence, vol. 43, no. 4, pp. 1369– 1379, 2019.
- [12] Y. Ouali, C. Hudelot, and M. Tami, "Semi-supervised semantic segmentation with cross-consistency training," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 12 674–12 684.
- [13] R. He, J. Yang, and X. Qi, "Re-distributing biased pseudo labels for semi-supervised semantic segmentation: A baseline investigation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 6930–6940.
- [14] L. Yang, W. Zhuo, L. Qi, Y. Shi, and Y. Gao, "St++: Make selftraining work better for semi-supervised semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4268–4277.
- [15] Y. Zou, Z. Zhang, H. Zhang, C.-L. Li, X. Bian, J.-B. Huang, and T. Pfister, "Pseudoseg: Designing pseudo labels for semantic segmentation," *arXiv preprint arXiv:2010.09713*, 2020.
- [16] X. Wei, X. Wei, X. Kong, S. Lu, W. Xing, and W. Lu, "Fmixcutmatch for semi-supervised deep learning," *Neural Networks*, vol. 133, pp. 166– 176, 2021.
- [17] X. Chen, Y. Yuan, G. Zeng, and J. Wang, "Semi-supervised semantic segmentation with cross pseudo supervision," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 2613–2622.
- [18] H. Chen, Y. Jin, G. Jin, C. Zhu, and E. Chen, "Semisupervised semantic segmentation by improving prediction confidence," *IEEE Transactions* on Neural Networks and Learning Systems, vol. 33, no. 9, pp. 4991– 5003, 2021.
- [19] I. Alonso, A. Sabater, D. Ferstl, L. Montesano, and A. C. Murillo, "Semisupervised semantic segmentation with pixel-level contrastive learning

from a class-wise memory bank," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 8219–8228.

- [20] Y. Zhong, B. Yuan, H. Wu, Z. Yuan, J. Peng, and Y.-X. Wang, "Pixel contrastive-consistent semi-supervised semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 7273–7282.
- [21] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill *et al.*, "On the opportunities and risks of foundation models," *arXiv preprint arXiv:2108.07258*, 2021.
- [22] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.
- [23] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig, "Scaling up visual and vision-language representation learning with noisy text supervision," in *International Conference on Machine Learning*. PMLR, 2021, pp. 4904–4916.
- [24] A. Shrivastava, A. Gupta, and R. Girshick, "Training region-based object detectors with online hard example mining," in *Proceedings of the IEEE* conference on computer vision and pattern recognition, 2016, pp. 761– 769.