SSR : SAM is a Strong Regularizer for domain adaptive semantic segmentation

Yanqi Ge△ Ye Huang△ Wen Li Lixin Duan□

Shenzhen Institute of Advanced Study, University of Electronic Science and Technology of China

***Abstract*—We introduced SSR, which utilizes SAM (segment-anything) as a strong regularizer during training, to greatly enhance the robustness of the image encoder for handling various domains. Specifically, given the fact that SAM is pre-trained with a massive-scale dataset that covers a diverse variety of domains, the feature encoding extracted by the SAM is obviously less dependent on specific domains when compared to the traditional ImageNet pre-trained image encoder. Meanwhile, the ImageNet pre-trained image encoder is still a mature choice of backbone for the semantic segmentation task, especially when the SAM is category-irrelevant. As a result, our SSR provides a simple yet highly effective design. It uses the ImageNet pre-trained image encoder as the backbone, and the intermediate feature of each stage (*i.e.* there are 4 stages in MiT-B5) is regularized by SAM during training. Extensive experiments show our SSR significantly improved performance over the baseline without introducing any extra inference overhead.**

***Index Terms*—semantic segmentation, domain adaption**

I. Introduction

Research of Unsupervised Domain Adaptation (UDA) aims to tackle the problem of domain gap. Current UDA methods typically focus on improving the model or using more advanced data augmentation strategies. Despite significant progress in recent years, their performance is still noticeably inferior to that of purely supervised models.

In this work, we were inspired by some of the latest ideas in the field, such as the foundation model and big data. We didn't want to limit ourselves to existing UDA patterns, so we decided to use big data to explore new ways of improving UDA performance. Due to budget constraints, we utilize the foundation model to benefit from internet data indirectly.

SAM (Segment-anything [1]) is a vision foundation model trained on a massive-scale dataset (SA-1B). SAM's feature map, though it cannot directly make pixel classification, is obviously more domain-independent than a traditional model trained on a small dataset. Thus, We proposed SSR (SAM is a Strong regularize), using the relatively robust SAM feature map as a regularizer during training to learn more domain-independent weights for UDA. During inference, because SAM is no longer required, the proposed SSR does not cause extra inference overhead. Experimentally, after the SAM intervention during training, our model's performance significantly improved compared to the baseline.

To summarize, our contributions are listed below:

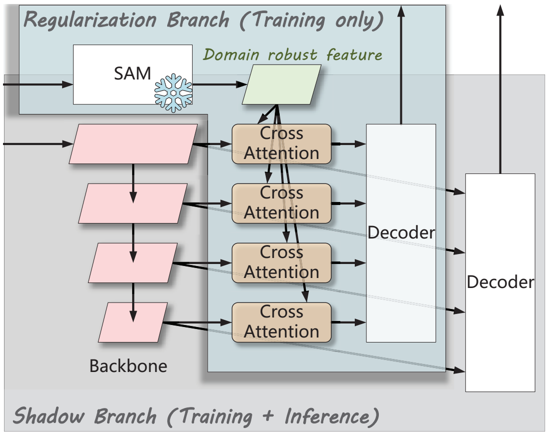

Fig. 1. The SSR architecture we proposed. Zoom in to see better.

- We utilize SAM to indirectly leverage big data during training and notably enhance UDA performance.
- We proposed 'shadow branch' to make SAM is not required for inference and avoids extra inference overhead.

TABLE I
Ablation studies of applying SAM regularization to different stages of the backbone feature map.

S0	S1	S2	S3	mIoU(%)	△
				68.3	-
			✓	68.8	0.5 ↑
		✓	✓	69.0	0.7 ↑
	✓	✓	✓	68.7	0.4 ↑
✓	✓	✓	✓	**69.3**	**1.0 ↑**

II. Proposed methods

Our proposed SSR is based on DAFormer, which utilizes MiT-B5 as its backbone and consists of two branches: a regularization branch for training only and a shadow branch for both training and inference.

A. Regularization branch

SSR froze the weights of SAM to make it under inference mode. The input image is through both SAM (Segment-Anything) and MiT-B5 [4]. SAM produces 1 × feature map (see green feature map in Fig. 1), while MiT-B5 produces 4 × feature maps (see pink feature maps in Fig. 1) from its pyramid structure, which consists of four stages.

As SAM is trained on a massive-scale dataset (SA-1B [1]) that covers diverse domains for each category, the feature map it generates is naturally less dependent on specific domains.

△ : Equal contributions, □ : Corresponding author

This research is supported by the Shenzhen LongHua Fundamental Research Program No. 10162A20230325B73A546 and Sichuan Science and Technology Program No. 2022YFG0189.

979-8-3503-5409-6/24/$31.00 ©2024 IEEE
DOI 10.1109/CAI59869.2024.00236

TABLE II

COMPARISON WITH STATE-OF-THE-ART METHODS FOR GTA5 → CITYSCAPES AND SYNTHIA → CITYSCAPES ADAPTATION TASKS.

	Road	S.walk	Build.	Wall	Fence	Pole	Tr.Light	Sign	Veget.	Terrain	Sky	Person	Rider	Car	Truck	Bus	Train	M.bike	Bike	mIoU(%)
									GTA5 → Cityscapes											
DAFormer [2]	95.7	70.2	89.4	53.5	48.1	49.6	55.8	59.4	89.9	47.9	92.5	72.2	44.7	92.3	74.5	78.2	65.1	55.9	61.8	68.3
SSR (DAFormer)	96.5	73.2	89.4	54.4	43.3	50.4	**56.5**	61.2	**90.0**	45.4	**92.8**	72.4	45.8	93.0	80.4	79.3	70.5	**58.4**	**64.7**	69.3
MIC [3]	96.9	76.2	**90.2**	**59.1**	50.7	**52.8**	55.6	60.5	**90.0**	**49.0**	**92.8**	**73.2**	**48.0**	92.6	75.8	**83.7**	69.4	54.1	62.7	70.1
SSR (MIC)	**97.3**	**78.2**	**90.2**	58.1	**51.9**	**52.8**	56.2	**64.6**	89.8	46.7	91.9	72.7	45.6	**92.9**	**82.5**	78.8	**72.0**	56.1	63.8	**70.6**
									Synthia → Cityscapes											
DAFormer [2]	84.5	40.7	88.4	41.5	6.5	50.0	55.0	54.6	86.0	–	89.8	73.2	48.2	**87.2**	–	53.2	–	53.9	**61.7**	60.9
SSR (DAFormer)	83.4	41.2	88.0	36.8	8.6	**52.5**	54.4	56.1	87.4	–	**94.0**	**75.1**	51.6	86.5	–	**66.8**	–	**58.4**	61.1	62.6
MIC [3]	83.0	40.9	88.2	37.6	9.0	52.4	**56.0**	**56.5**	**87.6**	–	93.4	74.2	51.4	87.1	–	59.6	–	57.9	61.2	62.2
SSR (MIC)	**89.0**	**55.7**	**89.3**	**47.3**	**10.5**	49.7	55.4	52.7	87.1	–	**94.0**	72.5	**49.8**	78.7	–	64.2	–	53.5	55.6	**62.8**

Hence, we directly use the SAM feature map to perform the cross-attention with 4 stages outputs of MiT-B5.

In each stage's cross-attention process, the proposed SSR utilizes the MiT-B5 backbone feature map as a query. To align the number of channels with the backbone feature, SSR applies a single linear layer to project the SAM feature map. Since there are four stages of outputs from the MiT-B5 backbone, SSR has four linear layers for channel projection.

After conducting the cross-attention computation, including the residual add, all of the outputs are passed to the decoder for the rest of the process.

During training, in order to minimize the loss, the cross-attention process ensures the backbone feature map has a similar distribution as the SAM feature map In simpler terms, it helps the model learn robust and adaptable weights that are not limited to a specific domain, which is crucial for accurate and effective domain adaptive semantic segmentation.

B. Shadow branch

In addition to the regularization branch, SSR also carries out a shadow branch that operates under the premise that the SAM only serves as a feature regularizer. The shadow branch is solely based on the segmentation model, which shares the backbone and decoder with the regularization branch and does not incorporate the SAM or cross-attention mechanism.

During inference, the shadow branch is the only branch that involves inference, resulting in zero extra inference overhead compared to the DAFormer baseline.

III. TRAINING DETAILS

We evaluate SSR based on MIC [3] and DAFormer [2] with a MiT-B5 encoder [4]. We follow the same training protocol with DAFormer and MIC. We use the pre-trained SAM (ViT-b) encoder to regularize the segmentation model.

IV. EXPERIMENTS

Ablation studies: We conducted ablation studies in Tab. I based on DAFormer, gradually adding SAM-regularized cross-attention to each backbone's stage. As shown in Tab. I, regularizing all stages resulted in the highest improvement.

Compare with baselines: We then compare our proposed SSR with DAFormer and MIC on GTA5→Cityscapes [5], [6] and Synthia→Cityscapes, respectively. As presented in Tab. II, both baselines exhibit improved performance after adding SSR, indicating that our proposed SSR can be generalized across multiple existing approaches.

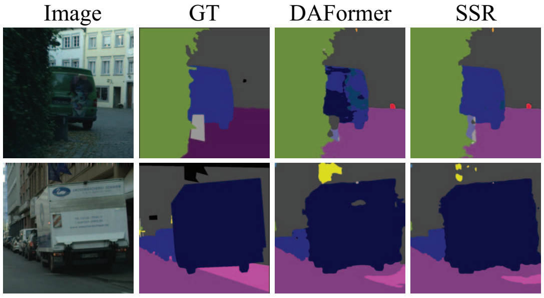

Fig. 2. Comparison of DAFormer vs SSR (DAFormer) on Cityscapes dataset.

Visualization: Fig. 2 provides visualizations for DAFormer and DAFormer + SSR. The visualization shows that using SAM as a regularizer during training reduces misclassification.

V. CONCLUSION

In this work, we introduce SSR (SAM is a Strong Regularizer), a new strategy to enhance the performance of domain adaptive semantic segmentation. SAM, which was trained on a massive-scale dataset, effectively regularizes the learning process of the model and enables it to produce more domain-independent representations. The shadow branch, which shares the same model, ensures that SSR has zero extra inference overhead. An ablation study and experiment are carried out to validate the effectiveness of the proposed SSR.

REFERENCES

[1] 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," in *ICCV*, 2023.

[2] L. Hoyer, D. Dai, and L. Van Gool, "Daformer: Improving network architectures and training strategies for domain-adaptive semantic segmentation," in *CVPR*, 2022.

[3] L. Hoyer, D. Dai, H. Wang, and L. Van Gool, "MIC: Masked image consistency for context-enhanced domain adaptation," in *CVPR*, 2023.

[4] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," in *NeurIPS*, 2021.

[5] S. R. Richter, V. Vineet, S. Roth, and V. Koltun, "Playing for data: Ground truth from computer games," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*. Springer, 2016, pp. 102–118.

[6] C. Marius, O. Mohamed, R. Sebastian, R. Timo, E. Markus, B. Rodrigo, F. Uwe, S. Roth, and S. Bernt, "The cityscapes dataset for semantic urban scene understanding," in *CVPR*, 2016.