BDSAM: Fusing Graph Attention Networks and Adapters to Address Domain Imbalance in Cross-Domain Segmentation

Yuefeng Zhao[1],Siting Zhou[1],Pengfei Sun[1],Junjie Wang[1],Haoyu Wang[1], Nannan Hu[1*]

[1]*Shandong Normal University, Jinan 25000, China*
{yuefengzhao,hunan246}@sdnu.edu.cn; {zhousiting0728,sunpengfei0519,ksh163853868}@163.com
1561663752@qq.com

***Abstract*—Cross-domain image segmentation plays a crucial role in the field of image segmentation, as it helps to improve the generalization ability of models across different datasets. However, existing models overemphasize domain-invariant features that are not specific to a particular domain, such as cells in medical images or buildings in remote sensing images. Consequently, they neglect the importance of domain-specific features, such as object contours or boundaries.Therefore, we propose Balanced Domain SAM(BDSAM), a model designed to balance domain-specific features while enhancing domain-invariant features.Specifically, we introduce an Iterative Optimization Selection Module (IOSM). By leveraging query features to continuously filter for more relevant pixels within the support features, we automatically select these pixels as point prompts. Furthermore, we propose a Domain-specific Feature Fusion Module (DFFM). By extracting and fusing features from different levels, we enhance the focus on domain-specific features.Extensive experiments on four cross-domain datasets show that our model has better average accuracy on 1-Shot and 5-Shot settings than the state-of-the-art approaches respectively .**

***Index Terms*—Cross-Domain Segmentation, Balance Domain, Graph Attention Networks,Adapter**

I. Introduction

Image segmentation, as one of the core tasks in modern computer vision, aims to precisely segment the regions of interest in an image to provide key semantic information for subsequent tasks. With the development of deep learning, segmentation models based on convolutional networks and self-attention mechanisms have achieved significant progress in modeling local details and global context, driving continual improvements in accuracy and robustness. However, the variability in data distributions and differences in acquisition devices in real-world scenarios still seriously constrain the generalization capability of models to new domains. Cross-domain image segmentation has become one of the hot topics in current research because it can be trained on a specific dataset and directly applied to another dataset with a different distribution.

* Corresponding Author.
This study was funded by the Shandong Provincial Natural Science Foundation Youth General Project (No.ZR2024QF060)(No.ZR2025QC1571), and National Natural Science Foundation of China (No.42271093)(No.62376034).

In recent years, significant progress has been made in this field [1] [2] [3].As a powerful and versatile visual segmentation model, the Segment Anything Model (SAM) [4] and its derivatives have achieved notable results in cross-domain segmentation applications [5] [6].UrbanCross utilizes SAM for key region extraction after segmentation and aligns these regions semantically with text to optimize visual-language matching [7].APSeg integrates support prototypes and pseudo query prototypes and transforms input features into a stable domain-invariant space [8].Despite achieving good segmentation results, these models tend to overemphasize domain-invariant features while neglecting domain-specific features, leading to overfitting to the style of the source domain.

In summary, we propose a method called Balanced Domain SAM(BDSAM).It balances domain-specific features while enhancing domain-invariant features to improve the model's generalization and accuracy. Specifically, we introduce an Iterative Optimization Selection Module (IOSM). Since the attention modules in the SAM mask decoder depend on the positional information encoded from point prompts, we use a graph attention network to aggregate similar features and automatically select these nodes as point prompts for the SAM decoder. In addition, to preserve domain-specific characteristics such as boundaries and textures, and to prevent the model from overfitting to the source-domain style, we propose a Domain-specific Feature Fusion Module (DFFM). This module uses adapters to extract features from different layers of the SAM decoder. After fusion, features from different layers are assigned different weights at each hierarchical level, thereby fully leveraging the multi-scale information provided by the SAM encoder.

Our main contributions are as following:

1.We propose a Balanced Domain SAM(BDSAM) network that enhances domain-invariant features while balancing domain-specific features.

2. We propose an Iterative Optimization Selection Module (IOSM) to enhance domain-invariant features. We also propose a Domain-specific Feature Fusion Module (DFFM) to extract and fuse features across different layers, achieving a balance between domain-invariant and domain-specific features.

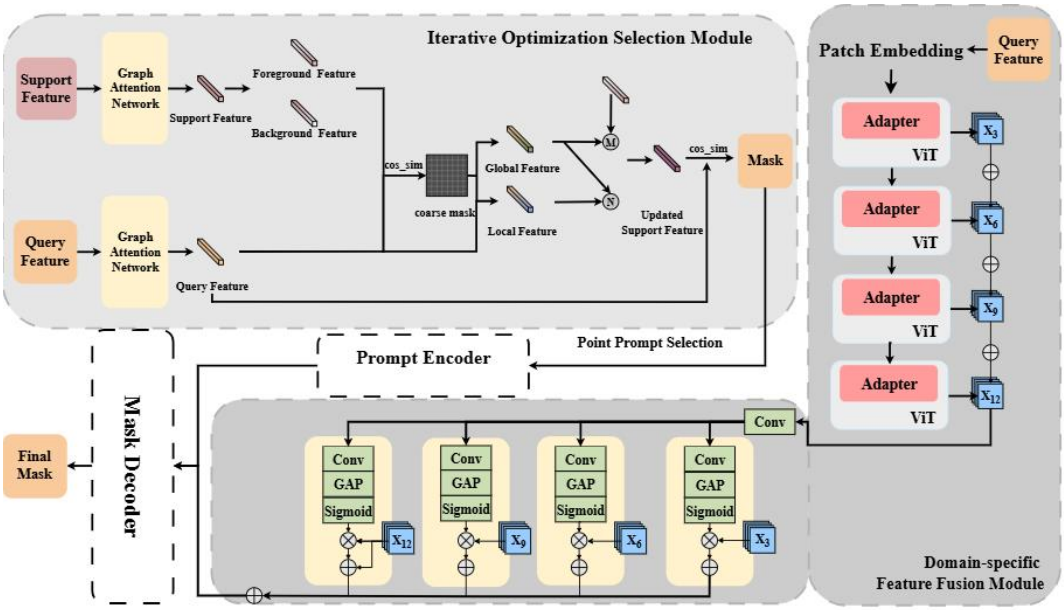


Fig. 1: Illustration of our framework, which integrates an iterative optimization selection module and a domain-specific feature fusion module.

3.In our extensive experiments across four cross-domain datasets, the mean-IoU achieved surpasses that of the current state-of-the-art methods.

II. Related works

A. SAM in cross-domain segmentation

In the field of artificial intelligence, the trend of large-scale foundation models in recent years has attracted widespread attention. As a powerful and general-purpose visual segmentation model, SAM [4] can generate masks for specific regions of interest through interactive clicks or the drawing of bounding boxes, thereby achieving significant results in cross-domain segmentation tasks. Common approaches in cross-domain segmentation include data augmentation, generating pseudo-labels for the target domain, and narrowing the domain gap through adversarial learning. Regarding pseudo-label generation, the LFMDA framework [9] proposes a SAM-guided local region homogeneity strategy to improve the quality of pseudo-labels. On the other hand, researchers from the Michigan team use geometry-aware progressive propagation to spread labels to all 3D points, thereby extending semantic segmentation from two dimensions to three. Concerning data augmentation, due to the drawbacks of low contrast and relatively poor signal-to-noise ratio in ultrasound images, the Nora framework [10], built upon the SAM structure, improves robustness by adding adaptive noise to target features. The MAUP framework [11] starts from the SAM prompt decoder and adaptively adjusts prompts according to the complexity of the target region to enhance the model's segmentation of target features. The UrbanCross architecture [7] further conducts adversarial training between the augmented text and image features to keep the feature distributions of the source and target domains consistent.

III. Our Method

Our method consists of two core components: the Iterative Optimization Selection Module (IOSM) and the Domain-Specific Feature Fusion Module (DFFM). The DFFM performs feature extraction and fusion through adapters. Our method is illustrate in Fig.1.

A. Iterative Optimization Selection Module

Due to differences between source domain and target domain datasets, the model requires domain-invariant feature extraction, retention, and enhancement. Therefore, we propose an iterative optimization selection module. Instead of manually annotating hints, it enhances efficiency and accuracy by computing and filtering pixels more relevant to the query features on the support features as point prompts.

First, we apply Graph Attention Networks (GAT) to the features. Specifically, for a given image I, the object masks $\{m_i\}_{i=1}^{l}$ generated by SAM, and the features F (F^s, F^q) extracted from the backbone, masked average pooling (MAP) is used to generate the initial visual prompts:

$$v_i = \text{MAP}(F, m_i). \tag{1}$$

Next, we calculate the importance between node i and node j.

$$w_{ij} = \frac{cos(v_i, v_j)}{\sum_{i=1}^{l} cos(v_i, v_j)}, \tag{2}$$

here, $cos(\cdot, \cdot)$ is the cosine similarity function, and w_{ij} represents the weight between visual prompts. The higher weight indicates stronger correlation between two nodes. Next, GAT aggregates the weighted average of all neighboring nodes of

node i and applies a residual connection to form the new feature representation of node i:

$$\bar{v}_i = \sum_{j=1}^{l} w_{ij} \cdot v_j + v_i. \tag{3}$$

We then use the inverse process of MAP, namely RMAP, to recover the visual prompts into feature maps:

$$\bar{F} = \mathrm{RMAP}(\bar{v}_i, m_i). \tag{4}$$

Both the query image and the support image undergo the same operation, resulting in $\bar{F}^{\,s}$ and $\bar{F}^{\,q}$.For each support image, foreground feature $\bar{F}^{\,s}_{fg}$ and background feature $\bar{F}^{\,s}_{bg}$ are obtained using the groundtruth mask. A coarse query mask is obtained by calculating the cosine similarity between the foreground (background) features and the support features:

$$M^1 = cos(\bar{F}^{\,s}, \bar{F}^{\,q}). \tag{5}$$

here, $M^1 = (M^1_{\mathrm{fg}}, M^1_{\mathrm{bg}})$, where M^1_{fg} and M^1_{bg} denote the coarse probability maps for the foreground and background respectively. Then, features with high confidence in these two regions are filtered out by setting a threshold τ.For global features, we can average the features within the foreground and background:

$$global_{\,fg} = \mathrm{MAP}\left(\bar{F}^{\,q},\, M^1_{\mathrm{fg}} > \tau_{fg}\right), \tag{6}$$

here, MAP represents Masked Average Pooling. The same operation can be used to obtain background global features $global_{\,bg}$. For each pixel location of the query feature, a normalized similarity with the foreground/background support features after filtering is used as a weight. The weighted average yields the local support features guided by the query feature:

$$local_{\,fg} = norm\_sim\left(\bar{F}^{\,q},\, M^1_{\mathrm{fg}} > \tau_{fg}\right). \tag{7}$$

The same operation can be used to obtain background local features $local_{\,bg}$. Compared to the stability of global features, local features are more detailed and can handle boundary details. After obtaining the global and local features from updates, we employ a linear fusion coefficient to fuse them with the initial support features, $\bar{F}^{\,s}_{fg}$and $\bar{F}^{\,s}_{bg}$. For the foreground:

$$\hat{F}^{\,s}_{fg} = 0.5 \cdot \bar{F}^{\,s}_{fg} + 0.5 \cdot global_{\,fg}, \tag{8}$$

$$\hat{F}^{\,s}_{bg} = 0.3 \cdot global_{\,bg} + 0.7 \cdot local_{\,bg}. \tag{9}$$

The final mask prediction $\bar{M}^{\,q}$ is generated by calculating the cosine similarity between the updated support feature $\hat{F}^{\,s}$ and the query feature $\bar{F}^{\,q}$:

$$\bar{M}^{\,q} = \mathrm{softmax}\big(cos(\bar{F}^{\,q},\, \hat{F}^{\,s})\big). \tag{10}$$

After obtaining the updated prediction mask $\bar{M}^{\,q}$, to make the prompts more representative.Inspired by maskSLIC [12], we generate a bounding box for the mask of the image, controlling the extraction of positive points from within the mask and negative points from outside the mask. This can guide SAM to enhance the segmentation of the target object and reduce background interference.

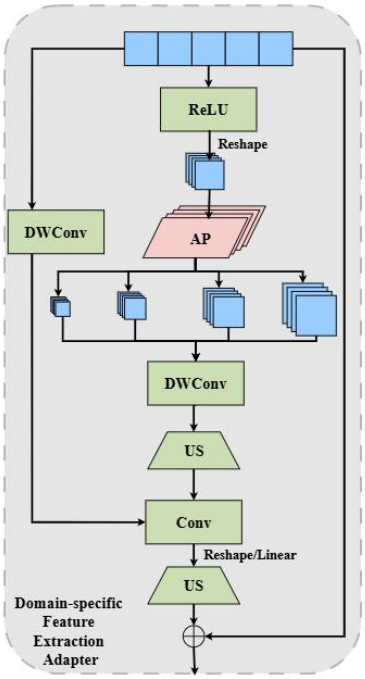


Fig. 2: Illustration of domain-specific feature extraction adapter.

B. Domain-specific Feature Fusion Module

In SAM encoder, the lower layers contain more domain-specific information, which is crucial for training the model's ability to extract edge details. Therefore, we insert an adapter before the first normalization layer of each ViT block in the SAM encoder, so as to extract both low-level and high-level features.

As shown in Fig.2, we first use a linear projection layer to reduce the feature dimension.After passing through a ReLU activation function, we reshape the features for further spatial information processing:

$$\mathbf{X}^r_i = Reshape\big(\mathrm{ReLU}(\mathbf{X}_i)\big). \tag{11}$$

Next, we employ four average pooling (AP) layers with different scales to obtain multi-resolution feature, capturing information from varying spatial ranges. Following this, we utilize 3x3 depthwise separable convolutional(DWConv) layers to enhance the local spatial structure representation of the features, ensuring that the SAM learns features such as edges and textures:

$$\mathbf{X}^s_{i,j} = \mathrm{US}\big(\mathrm{DWConv}\big(\mathrm{AP}(\mathbf{X}^r_i)\big)\big), \quad 1 \le j \le 4. \tag{12}$$

Spatial features from different scales are concatenated along the channel dimension and then fused using 1×1 convolution to form richer content features:

$$\bar{\mathbf{X}}_i = \phi_{1\times1}\Big(\big[\mathbf{X}^s_{i,j},\, \mathrm{DWConv}(\mathbf{X}^r_i)\big]\Big) \tag{13}$$

Finally, we reshape the features into tokens, then use a Linear layer to up-project them back to the original dimension,

TABLE I: Comparison of model performance in 1-shot and 5-shot settings. The best methods are highlighted in **bold** respectively.

Methods	Backbone	Publication	Deepglobe		ISIC		Chest X-ray		FSS-1000		Average	
			1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
PATNet [13]	Res-50	ECCV-22	37.9	43.0	41.2	53.6	66.6	70.2	78.6	81.2	56.07	62.00
ABCDFSS [14]	Res-50	CVPR-24	42.6	49.0	45.7	53.3	79.8	81.4	74.6	76.2	60.68	64.98
DRA [15]	Res-50	CVPR-24	41.3	50.1	40.8	48.9	82.4	82.3	79.1	80.4	60.90	65.43
APSeg [8]	ViT	CVPR-24	35.9	40.0	45.4	54.0	84.1	84.5	79.7	81.9	61.28	65.10
AFM [16]	Res-50	NIPS-24	40.9	44.9	41.7	51.2	78.2	82.8	79.3	81.8	60.03	65.18
IFA [17]	Res-50	CVPR-24	50.6	58.8	66.3	69.8	74.0	74.6	80.1	82.4	67.75	71.40
SDRC [18]	ViT	ICML-25	43.2	46.8	46.6	55.0	82.9	84.8	80.3	82.6	63.40	67.30
BDSAM	Res-50	Ours	**53.6**	**61.1**	**69.8**	**75.2**	**84.5**	**86.4**	**83.1**	**84.9**	**72.75**	**76.90**

and apply a residual connection with the input tokens to obtain the final output:

$$\hat{X}_i = \text{Linear}\big(Reshape(\bar{\mathbf{X}}_i)\big) + \mathbf{X}_i. \tag{14}$$

After extracting the multi-scale information, if only the final output of the last layer of the image encoder is used as input to the mask decoder, the multi-scale information cannot be integrated. Therefore, we assign fusion features to different levels using different weight assignments. We extract output features from the 3rd, 6th, 9th, and 12th layers of the SAM encoder and denote them as $\hat{X}_i (i = 3, 6, 9, 12)$. First, we concatenate them through 1×1 convolutions to obtain the aggregated feature:

$$\hat{X}_c = \phi_{1\times1}\big([\hat{X}_3,\, \hat{X}_6,\, \hat{X}_9,\, \hat{X}_{12}]\big). \tag{15}$$

Then, we use Global Average Pooling (GAP) to transform each feature map into a single numerical value, representing the overall response of that feature map. Subsequently, a 1×1 convolution is used to generate weightsθ .

$$\theta = \delta\big(\text{GAP}(\phi_{1\times1}(\hat{X}_c))\big), \tag{16}$$

where δ denotes the Sigmoid function.Finally, we obtain the fused feature F_f as follows:

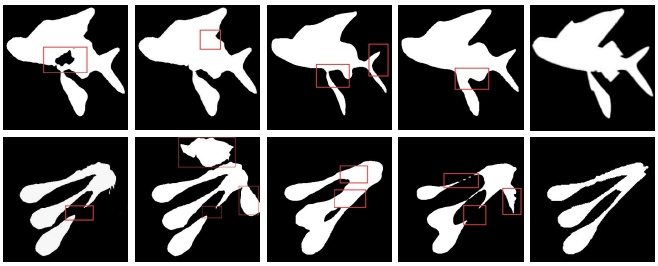
$$F_f = \sum_{i=3,6,9,12} (\theta \cdot \hat{X}_i + \hat{X}_i). \tag{17}$$

IV. EXPERIMENTS

A. Datasets and Implementation Details

We train our model on the PASCAL VOC 2012 [19] dataset augmented with SBD [20], then evaluate the trained model on four target domain datasets: Deepglobe [21], ISIC [22], Chest X-Ray [23], and FSS-1000 [24].

Deepglobe is a satellite remote sensing dataset includes urban areas, agricultural land, shrub-grassland/wooded meadows, forest, water bodies, wasteland, and unknown classes., and unknown. **ISIC** is a medical image dataset for skin cancer screening, containing 2594 skin lesion images including three major types of skin lesions. **Chest X-ray** is a dataset of tuberculosis diagnosis X-ray images, containing 566 medical images . **FSS-1000** is a natural image dataset designed for few-shot segmentation, containing 1000 target categories, with 10 images annotated per category.



PATNet [13] ABCDFSS [14] APSeg [8] IFA [17] BDSAM

Fig. 3: Visual comparison of our method and previous CD-FSS methods.

Our baseline model is SSP [25], using a ResNet-50 pre-trained on ImageNet [26] as the backbone. For SAM, we use the base version. All images are resized to 400 × 400 for input. The numbers of positive and negative point prompts are both set to 20. During training, we set the batch size to 6 and the number of episodes to 6000. The learning rate starts at 0.001 and is gradually reduced to 0.0001. We use SGD optimizer with momentum 0.9 and weight decay 5e-4.For evaluation, we use mean Intersection over Union (mIoU) as the evaluation metric, and report average results over 5 random seeds. Each run consists of 600 episodes sampled from Deepglobe, ISIC, and Chest X-ray respectively, and 1200 episodes from FSS-1000.

B. Comparisons with State-of-the-arts

Table I shows a comparison between BDSAM and the previous CD-FSS methods under 1-shot and 5-shot settings. In the 1-shot setting, our method surpasses the best method by 5% in average IoU, and in the 5-shot setting by 5.5%. This establishes that our proposed method can achieve strong performance in CD-FSS tasks. Furthermore, compared with other SAM-based methods (such as APSeg, He et al., 2024) in 1-shot and 5-shot settings, GPRN achieves significant improvements of 14.47% and 11.8%, respectively. These results highlight the effectiveness of the proposed module in fully leveraging SAM's potential in CD-FSS.

Fig. 3 presents visual results compared with other methods. We used the same image and compared five methods: PATNet

TABLE II: Effects of each proposed component on the Deepglobe dataset.

Method	1-shot	5-shot
baseline	48.8	56.5
baseline +IOSM	52.4	59.2
baseline +DFFM	52.1	58.9
baseline +IOSM+DFFM	**53.6**	**61.1**

TABLE III: Effects of number of point prompts on Deepglobe dataset.

Foreground	Background	1-shot mIoU
10	0	50.4
10	10	52.5
20	20	**53.6**
30	30	52.8

[13], ABCDFSS [14], APSeg [8], IFA [17], and BDSAM. The last column shows the results obtained by our model. It is not hard to see that the segmentation results from the remaining methods will, to some extent, have missing parts, especially at the edges of objects or at regions where objects are stacked. Our method can not only identify the target object, but also make the segmentation at the edge parts more apparent. This reflects the advanced nature of our method.

C. Ablation Studies

1) Effects of each proposed component: Table II shows the impact of key components of BDSAM on the accuracy of the Deepglobe dataset. The first line represents our baseline. It can be seen that in the 1-shot setting, simply adding IOSM, which enhances domain-invariant features by iteratively mining reliable point prompts, increases mIoU by 3.6%. Likewise, simply adding DFFM, which uses the SAM encoder to extract and fuse multi-scale features to strengthen domain-specific features, increases mIoU by 3.3%. When both are combined, the improvement on top of the baseline reaches 4.8%. This supports our "balancing" claim: they contribute from different perspectives and achieve the best results when used together.

Fig. 4 shows the segmentation results obtained by our components. From left to right, they are the image, ground truth, the IOSM segmentation map, the DFFM segmentation map, and the BDSAM segmentation map. It can be seen that whether only IOSM is added or only DFFM is added, the segmentation results are not clear. For example, in the second row, in experiments on the Deepglobe dataset under 1-shot setting, simply adding IOSM can identify the water bodies in the image, but it also causes other categories, such as urban areas and agriculture, to be treated as target regions. When only adding DFFM, the background and foreground become confused, making it impossible to correctly segment the target regions.Only when the two modules are used together can the segmentation achieve the best performance.

2) Effects of number of point prompts: The number of point prompts affects the segmentation of the target object and the interference from the background. From the first and second rows of Table III, it can be seen that selecting negative point prompts from the background can also improve performance. When the number of point prompts for both foreground and background is set to 20, the model achieves the best performance.However, when both are set to 30, the mIOU decreases by 0.8% and the computation increases. Therefore, we ultimately set the number of point prompts for both foreground and background to 20.

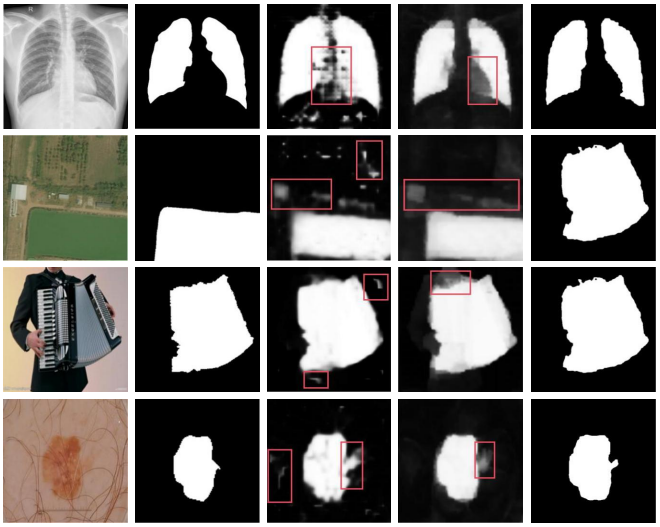


Fig. 4: Visualization of each proposed component segmentation.

TABLE IV: Effects of different attention mechanisms on the Deepglobe dataset.

Method	1-shot	5-shot
Self-attention	51.8	57.8
Multi-head attention	51.6	59.2
Spatial attention	52.5	60.3
Graph attention	**53.6**	**61.1**

3) Effects of different attention mechanisms.: In IOSM, we use graph attention networks to enhance image feature representations. In addition, we compare with several other ways of enhancing feature representations. On the Deepglobe dataset, we operate on features using self-attention networks, multi-head attention networks, spatial attention networks, and graph attention networks, respectively. As shown in the TABLE IV, applying graph attention to the features yields the best mIOU, improving by about 1%-3% over the other attention mechanisms. This indicates that, for the cross-domain image segmentation task, graph attention networks can effectively aggregate information of the segmentation target, thereby boosting expressive power.

4) Effects of feature fusion at different levels in the adapter.: In the DFFM module, we separately extract the low-level and high-level features of each ViT from the SAM encoder

TABLE V: Effects of feature fusion at different levels in the adapter on the Deepglobe dataset

Method	1-shot
Domain-invariant	48.9
Domain-specific	50.1
Both	53.6

and fuse them. We conducted a comparative study on the Deepglobe dataset. The TABLE V shows that if only domain-specific features are extracted and fused, the mIoU decreases by 3.5%. Likewise, if only domain-invariant features are extracted and fused, the mIoU decreases by 4.7%. Only when both are extracted and fused together does the performance reach its best. This demonstrates the importance of balancing domain-invariant features and domain-specific features in cross-domain segmentation, and also highlights the necessity of our research.

V. CONCLUSION

In this work, we propose a Balanced Domain SAM module(BDSAM) to address the imbalance of attention to domain-specific and domain-invariant features in cross-domain image segmentation tasks. Specifically, we introduce an Iterative Optimization Selection Module (IOSM). It continuously searches for more relevant and reliable pixels in support features using query features and automatically selects these as point prompts. Furthermore, to preserve domain-invariant features and prevent model overfitting, we propose a Domain-Specific Feature Fusion Module (DFFM). This module extracts and fuses features from different layers of the SAM decoder, then attaches them to features at different scales with varying weights.The model we propose demonstrates state-of-the-art performance on all four cross-domain datasets.

REFERENCES

[1] H. Basak and Z. Yin, "Semidavil: Semi-supervised domain adaptation with vision-language guidance for semantic segmentation," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 9816–9828.

[2] Y. Tian, C. Wen, M. Shi, M. M. Afzal, H. Huang, M. O. Khan, Y. Luo, Y. Fang, and M. Wang, "Fairdomain: Achieving fairness in cross-domain medical image segmentation and classification," in *European Conference on Computer Vision*. Springer, 2024, pp. 251–271.

[3] L. He and S. Todorovic, "Attention decomposition for cross-domain semantic segmentation," in *European Conference on Computer Vision*. Springer, 2024, pp. 414–431.

[4] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 4015–4026.

[5] H. Sun, R. Gong, I. Nejjar, and O. Fink, "Dynalign: Unsupervised dynamic taxonomy alignment for cross-domain segmentation," *arXiv preprint arXiv:2501.16410*, 2025.

[6] Z. Cheng, J. Guo, J. Zhang, L. Qi, L. Zhou, Y. Shi, and Y. Gao, "Mamba-sea: A mamba-based framework with global-to-local sequence augmentation for generalizable medical image segmentation," *IEEE Transactions on Medical Imaging*, 2025.

[7] S. Zhong, X. Hao, Y. Yan, Y. Zhang, Y. Song, and Y. Liang, "Urbancross: Enhancing satellite image-text retrieval with cross-domain adaptation," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 6307–6315.

[8] W. He, Y. Zhang, W. Zhuo, L. Shen, J. Yang, S. Deng, and L. Sun, "Apseg: Auto-prompt network for cross-domain few-shot semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 23 762–23 772.

[9] W. Liu, P. Duan, Z. Xie, X. Kang, and S. Li, "Learning from vision foundation models for cross-domain remote sensing image segmentation," *IEEE Transactions on Image Processing*, 2025.

[10] Z. Wei, C. Wu, H. Du, R. Yu, B. Du, and Y. Xu, "Noise-robust tuning of sam for domain generalized ultrasound image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2025, pp. 476–486.

[11] Y. Zhu and H. Zhang, "Maup: Training-free multi-center adaptive uncertainty-aware prompting for cross-domain few-shot medical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2025, pp. 326–336.

[12] B. Irving, "maskslic: regional superpixel generation with application to local pathology characterisation in medical images," *arXiv preprint arXiv:1606.09518*, 2016.

[13] S. Lei, X. Zhang, J. He, F. Chen, B. Du, and C.-T. Lu, "Cross-domain few-shot semantic segmentation," in *European conference on computer vision*. Springer, 2022, pp. 73–90.

[14] J. Herzog, "Adapt before comparison: A new perspective on cross-domain few-shot segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 23 605–23 615.

[15] J. Su, Q. Fan, W. Pei, G. Lu, and F. Chen, "Domain-rectifying adapter for cross-domain few-shot segmentation," in *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, 2024, pp. 24 036–24 045.

[16] J. Tong, Y. Zou, Y. Li, and R. Li, "Lightweight frequency masker for cross-domain few-shot semantic segmentation," *Advances in Neural Information Processing Systems*, vol. 37, pp. 96 728–96 749, 2024.

[17] J. Nie, Y. Xing, G. Zhang, P. Yan, A. Xiao, Y.-P. Tan, A. C. Kot, and S. Lu, "Cross-domain few-shot segmentation via iterative support-query correspondence mining," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 3380–3390.

[18] J. Tong, Y. Zou, G. Chen, Y. Li, and R. Li, "Self-disentanglement and re-composition for cross-domain few-shot segmentation," *arXiv preprint arXiv:2506.02677*, 2025.

[19] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International journal of computer vision*, vol. 88, no. 2, pp. 303–338, 2010.

[20] B. Hariharan, P. Arbeláez, L. Bourdev, S. Maji, and J. Malik, "Semantic contours from inverse detectors," in *2011 international conference on computer vision*. IEEE, 2011, pp. 991–998.

[21] I. Demir, K. Koperski, D. Lindenbaum, G. Pang, J. Huang, S. Basu, F. Hughes, D. Tuia, and R. D. Raskar, "A challenge to parse the earth through satellite images. arxiv 2018," *arXiv preprint arXiv:1805.06561*, 2018.

[22] N. Codella, V. Rotemberg, P. Tschandl, M. Celebi, S. Dusza, and D. Gutman, "a challenge hosted by the international skin imaging collaboration (isic)," 2018.

[23] S. Candemir, S. Jaeger, K. Palaniappan, J. P. Musco, R. K. Singh, Z. Xue, A. Karargyris, S. Antani, G. Thoma, and C. J. McDonald, "Lung segmentation in chest radiographs using anatomical atlases with nonrigid registration," *IEEE transactions on medical imaging*, vol. 33, no. 2, pp. 577–590, 2013.

[24] X. Li, T. Wei, Y. P. Chen, Y.-W. Tai, and C.-K. Tang, "Fss-1000: A 1000-class dataset for few-shot segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2869–2878.

[25] Q. Fan, W. Pei, Y.-W. Tai, and C.-K. Tang, "Self-support few-shot semantic segmentation," in *European conference on computer vision*. Springer, 2022, pp. 701–719.

[26] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.