MADA: Mask Aware Domain Adaptation for Open-set Semantic Segmentation

Keying Zhang

China Southern Power Grid Digital Power Grid Group Co., Ltd., Guangzhou, China
zhangky@csg.cn

Abstract

Open-set semantic segmentation aims to identify categories that extend beyond the scope of the training data. Compared to the conventional semantic segmentation, open-set semantic segmentation constitutes a more practical and challenging scenario. Nonetheless, prevalent open-set semantic segmentation models predominantly incorporate extensive image-text datasets and substantial network architectures. Although the design enhances the comprehensive performance of these models, it also intensifies their computational demand, making them considerably challenging to train or fine-tune for adaptation to task-specific applications or domains. In this paper, we introduce a novel strategy called Mask Aware Domain Adaptation (MADA) for addressing open-set semantic segmentation challenges. MADA investigates the similarities between visual and text modalities in both the source and target domains, aiming to align all modalities efficiently with few data and computational resources. This alignment significantly enhances model performance in the target domain while simultaneously maintaining open-set capacity. Extensive experiments demonstrate the effectiveness and efficiency of our approach. We consider MADA to be a practical solution for scenarios which require high target domain performance as well as open-set flexibility capacity.

Introduction

Semantic segmentation classifies each pixel of a given image with a specific semantic category or class. Open-set or Open-vocabulary (Geng, Huang, and Chen 2020) Compared to conventional semantic segmentation, open-set semantic segmentation is a more challenging extension where the set of possible semantic classes or labels is not predefined or limited in the training phase. The open-set scenario offers flexibility to machine learning models in handling unseen inputs, enhancing adaptability and performance. It improves model robustness and stability in real-world applications. Previously, open-set scenarios were tackled through zero-shot methodologies, using visual-text correlations to align visual and language domains. However, this approach is limited by dataset quality and alignment to target applications. Generally, the overall performance remains relatively low for practical applications due to both the dataset and network capacity limitations (Xian et al. 2019; Bucher et al. 2019; Zhang et al. 2024).

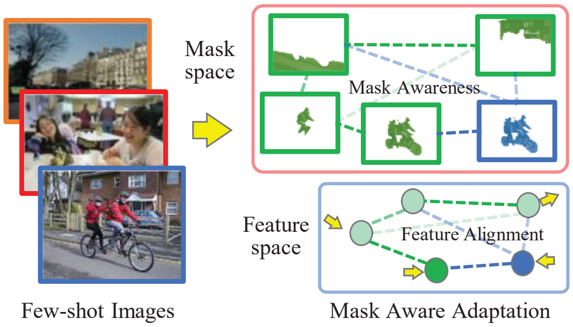

Figure 1: Concept of our Mask Aware Domain Adaptation strategy. Mask proposals associated with the corresponding features could be obtained from a pre-trained model. Then, our approach aims to efficiently align mask proposals with corresponding features from the target domain with significantly limited data and computational requirements. The basic assumption is that similar masks in visual space should share similar representations in feature space.

In recent years, there have been significant improvements in large language models (LLM) (e.g., BERT (Devlin et al. 2019) and GPT (Brown et al. 2020)) and foundation vision models (e.g., CLIP (Radford et al. 2021) and SAM (Kirillov et al. 2023)), which demonstrate impressive performance on open-set related tasks by mapping class specific text queries to image content. CLIP is a representative work which is trained on web-scale image-caption pairs. It effectively projects both visual and text signals to the same feature space. X-Decode(Zou et al. 2023; Liu et al. 2023) and Grounding DINO(Liu et al. 2023) integrate the LLM as well as the foundation models to solve the pixel-level segmentation problem. However, these approaches require a large amount of computational resources in the training and even fine-tuning phrases. For instance, one of the CLIP model backbones (i.e., RN50x64) takes 592 V100 GPUs and 18 days to complete the training process. This scale of resource requirement is not feasible for most budget-sensitive and task-specific applications.

Adapter-based parameter-efficient fine-tuning is a prac-

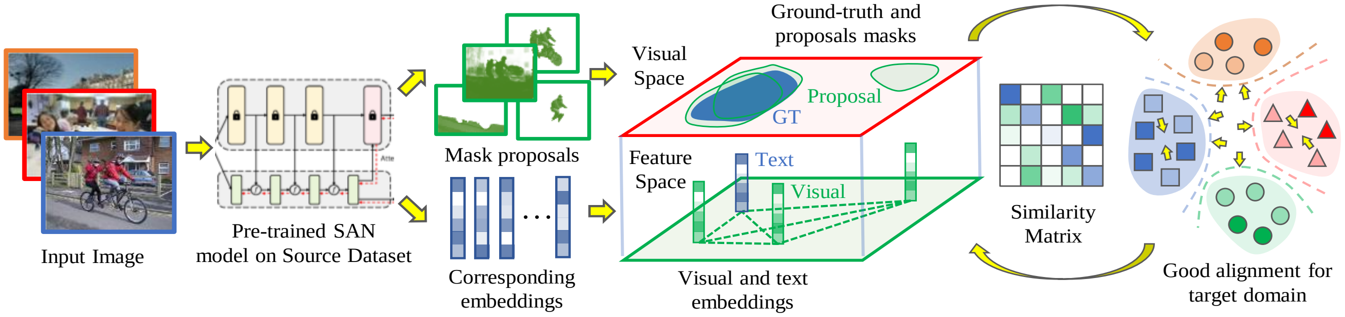
Input Image
Pre-trained SAN model on Source Dataset
Mask proposals
Corresponding embeddings
Visual Space
Feature Space
Ground-truth and proposals masks
Proposal
GT
Text
Visual
Visual and text embeddings
Similarity Matrix
Good alignment for target domain

Figure 2: Framework of our MADA approach. Image samples are set as inputs to raw segmentation models (e.g., SAN). Then, both the visual masks as well as the corresponding feature representations are obtained. We assume the more similar of the predicted masks compared with the ground truth, the more similar the feature representation is. To this end, MADA jointly explores the similarities of masks and representations, aligning the shift in both source and target domains.

tical solution which utilizes a relatively small-scale trainable network and a general size of training data to achieve feasible model fine-tuning. In a semantic segmentation scenario, Side Adapter Network (SAN) (Xu et al. 2023) provides a lightweight network structure which "attached" to a frozen CLIP model. A semantic segmentation dataset (e.g., COCO) is used to train and enable the segmentation capacity, while the natural CLIP model further enables the open-set capacity. SAN only requires a single GPU running for several hours which considerably reduces the resource requirements. Although SAN significantly reduces training intensity, there are still thousands of high-quality samples required to achieve elegant performance, which is still hard for some data-sensitive applications.

In this paper, we proposed a Mask Aware Domain Adaptation (MADA) approach for an open-set semantic segmentation scenario, the concept of MADA is shown in Figure 1. Specifically, given limited target domain samples (e.g., 50 samples), MADA fully explores the correlations of the mask proposals as well as the ground-truth masks, a similar strategy is also conducted in feature space. All these correlations are jointly optimized to obtain a simple but effective projection and directly deployed to raw models. Specific designed constraints are further proposed which preserves the open-set capacity of the raw model. Figure 2 illustrates the framework of the MADA model. In this way, we are able to use remarkably few samples to considerably improve the overall performance in the target domain. The main contributions of this work are listed below:

- A Mask Aware module is proposed, which jointly considers the mask similarity in visual space as well as the feature similarity in feature space.
- Structural consistency constraints are proposed which balance the target-specific focuses and the generalization for open-set scenarios.
- An efficient optimization solution is further derived which vastly reduces the computational cost to obtain the final results.

Extensive experiments demonstrate the effectiveness and efficiency of our MADA approach, which shows its feasibility for practical deployments in a wide range of applications.

Related work

Domain Adaptation

In many real-world scenarios, it is often expensive or impractical to collect large-scale labelled data for task-specific applications. The goal of domain adaptation is to develop methods that can leverage knowledge from a source domain where labelled data is abundant and transfer this knowledge to a target domain with limited or no labelled data. The challenges in domain adaptation arise from the differences in the distribution of the data between the source and target domains. The key is to explore methods that can effectively align the source and target domains to mitigate the negative impact of distribution shifts on model performance. Various strategies have been proposed to solve these challenges. For unsupervised domain adaption scenarios where only samples from the target domain are available without labels, (Tzeng et al. 2017) employs adversarial strategies to align domain features, where an adversarial network is used to distinguish between source and target domains. For semi-supervised domain adaption, (Yan et al. 2022) regularizes consistency across different views of target domain samples by aligning domains using a prototype-based optimal transport method, enhancing feature representations with a classwise contrastive clustering loss, and improving prediction accuracy through consistency-based self-training. (Zhao et al. 2021) enhances the source-target per-class separation before domain adaptive feature embedding learning to alleviate the negative effect of domain alignment on few-shot learning. More related interesting works include zero-shot domain adaption(Jhoo and Heo 2021; Lengyel et al. 2021), one-shot domain adaption(Luo et al. 2020), and open-set domain adaption(Jing, Liu, and Ding 2021; Bucci, Loghmani, and Tommasi 2020).

However, most existing studies either assume that most source categories and the target categories are expected to be identical, or there should be no unseen categories in the testing phase. This setting makes it them hard to conduct domain adaptation while still preserving generalize on open-set scenarios.

Open-set Semantic Segmentation

Semantic segmentation classifies each pixel of a given image with a specific label. In traditional (i.e., close-set) semantic segmentation, models are trained based on the fixed set of classes. However, in open-vocabulary semantic segmentation, the model is not restricted to a predefined set of classes and is capable of discovering and assigning labels to regions that may not have been seen during training. Open-set or Open-vocabulary semantic segmentation is a more practical and challenging extension. More advanced algorithms and powerful models in both language and visual domains (Fu et al. 2023) are required to achieve this goal. ZS3Net (Bucher et al. 2019) combines a segmentation model with a method to generate visual representations from semantic word embeddings. ZS3Net could handle pixel classification tasks of both seen and unseen categories. X-Decode (Zou et al. 2023) proposed an advanced decoder which decodes pixel-level classification results via arbitrary visual and query inputs. Grounding-DINO (Liu et al. 2023) performs visual-language modality fusion at multiple stages, including feature enhancement, language-guided query selection, and cross-modal decoders on object detection tasks. Grounding-SAM combines the capacity of DINO with the high precision segmentation model, SAM (Kirillov et al. 2023), to achieve higher performance. DINO-V2 (Oquab et al. 2023) explores multiple existing approaches to scale the pre-training in terms of data and model size for more robust visual feature learning. CLIP-(Radford et al. 2021) learns the visual-language correlations from web-scale visual-text pair training data. CLIP is able to project both visual and text data into the same feature space. SAN(Xu et al. 2023) utilizes a pre-trained vision-language model and treats semantic segmentation as a region recognition problem. By attaching a side network to a frozen CLIP model, the SAN framework enables the reuse of features.

Although these models achieve elegant performance for open-set semantic segmentation, all the models still require considerable high-quality training data for either training or fine-tuning which cannot effectively and efficiently adapt to task-specific applications with limited resources.

Our method

Our MADA approach contains two main modules associated with an effective and efficient optimization solution. The details are introduced below:

Mask Aware Correlation Learning

Given an image $I \in \mathbb{R}^{h \times w \times 3}$ to raw segmentation model, we can obtain the mask proposal $M \in \mathbb{R}^{h \times w \times n_m}$ as well as the corresponding visual feature vectors $f_I \in \mathbb{R}^{d \times n_m}$ of the proposed masks. h and w are the image height and width, n_m is the number of mask proposals, d is the dimension of the embedded visual feature. Assume N images are used in our learning phrase, by concatenating all the embedded visual feature together, we can obtain the visual feature matrix $F \in \mathbb{R}^{d \times n_f}$, where $n_f = n_m \cdot N$. In our approach, we aim to align the feature as well as the text space as efficient as possible. To this end, a liner project $P \in \mathbb{R}^{d \times d}$ is proposed, which is used to project embeddings into a modified space. The equations are shown below:

$$\hat{F} = PF, \tag{1}$$

where $\hat{F}$ is the projected features.

To obtain the similarity scores between visual and text embeddings, we further infer the text encoding matrix, $E \in \mathbb{R}^{d \times n_l}$ from the target label list, where n_l is the label number of the target dataset. Since CLIP model is able to project both visual and text data input the same feature space, to this end, E and F share the same feature space as well as the dimension number d. Thus, the cosine similarity could be obtained by the equation below:

$$s_{ij} = \frac{f_i^\top \cdot e_j}{\|f_i\| \cdot \|e_j\|}, \tag{2}$$

where $f_i \in \mathbb{R}^d$, $e_j \in \mathbb{R}^d$ denote the i-th and j-th feature vectors from F and E. $\|f_i\|$ and $\|e_j\|$ are their corresponding norms, and s_{ij} is the cosine similarity of f_i and e_j. By extending Eq.(2) to a matrix format, we can have:

$$S = \hat{F}E^\top = PFE^\top, \tag{3}$$

where $S \in \mathbb{R}^{n_l \times n_f}$ are the similarity score matrix which denotes the similarity of each pair-wise visual feature as well as text encoding.

As introduced in the above section, not all visual and textual features are related. Our principle assumption is that the similarities of the mask proposal compared with ground-truth is, the similar these feature representations supposed to be. More specifically, the basic goal of this module is to align the features with high correlation proposals while ignoring the other mismatched proposals. To achieve this, obtaining the similarity of two pair-wise visual masks is crucial. The basic idea is an IoU (Intersection over Union)-like strategy which is calculated by dividing the area of intersection between two masks by the area of their union. The value ranges from 0 to 1 which indicates from no overlap to perfect overlap. However, we consider this strategy cannot fully reflect the correlations between visual and textural domains. First, the sizes of neither mask proposals nor the ground-truth masks are ignored, while this information is important for aligning the intensity level of the visual features. Second, due to the sparsity of the mask categories in each image sample, assigning the 0 weights could easily cause unpredictable optimization results.

To this end, our mask-aware approach includes both the IoU-like calculation as well as size and non-label situation together. The overall equation is shown below:

$$c = \begin{cases} \frac{m_{\mathrm{s}_{pred}} \cap m_{gt}}{m_{\mathrm{s}_{pred}} \cup m_{gt}} e^{\alpha \frac{s_{pred}}{s\mathrm{s}_{image}}}, & m_{\mathrm{s}_{pred}} \cap m_{gt} > 0, \\ -\lambda \cdot e^{\alpha \frac{s\mathrm{s}_{pred}}{s\mathrm{s}_{image}}}, & m_{\mathrm{s}_{pred}} \cap m_{gt} = 0, \end{cases} \tag{4}$$

where c is the mask-aware of the correlation score of a given mask proposal m_{pred} and the ground-truth mask m_{gt}, s_{pred} and s_{image} are the corresponding sizes, α and λ are the trade-off parameter which tunes the scales of each term. More specifically, when there are overlaps compared with ground truth, both the IoU-like and size ratio will be used to tune the contribution to overall correlation scores. If there are no overlaps with ground-truth masks, a minor penalty weight would be involved to stabilize the learning procedure.

Given $c_i \in \mathbb{R}^{n_l}$ which is the correlation vector of a single mask, and $s_j \in \mathbb{R}^{n_l}$ is similarity score in feature space, we could obtain the overall weight by sum them up:

$$v_{ij} = c_i^\top s_j, \tag{5}$$

To get the overall scores from the entire samples, we sum all v_{ij} together and our goal is to optimize P which makes the overall similarity scores are greater as possible. The equation is shown as below:

$$\max_P \text{Tr}(CE^\top PF), \tag{6}$$

where $\text{Tr}(\cdot)$ denotes the trace of a matrix, which is the sum of elements on the main diagonal. $C \in \mathbb{R}^{n_l \times n_f}$ is the mask correlation matrix where each element c_{ij} denotes the similarity scores of all the training samples.

Structure Consistent Mapping

We consider Eq.(6) is the major objective of our approach. However, it has a naïve solution which is assigning P are greater as possible. In addition, due to the sparsity of the ground-truth masks associated with the limited number of training samples in our design, it is easily cause overfitting if we only solve Eq.(6). Moreover, we still want to preserve the model generalization for open-set performance. To this end, a few structure consistent constraints are proposed which are deployed to illuminate above mentioned issues. Our final objective which is shown below:

$$\max_P \text{Tr}(CE^\top PF) - \mu_1 \|P - I\|_{\text{F}}^2 - \mu_2 \|P\|_{\text{F}}^2, \tag{7}$$

where $I \in \mathbb{R}^{d \times d}$ is an identity matrix, $\|\cdot\|_{\text{F}}^2$ is the 2-norm. $\|P - I\|_{\text{F}}^2$ limits the variations of P compared with the identity matrix, which is able to control the structural consistency and improve the generalization flexibility for other open-set inputs. $\|P\|_{\text{F}}^2$ controls the overall scales of projection P. μ_1 and μ_2 are trade-off parameters which balance the final objective and the constraints.

Optimization

Deploying iterative optimization (e.g., backpropagation) is a popular solution. However, it requires extra computational cost and the iteration results are highly dependent on the learning parameters (e.g., learning rate and batch size) which is not the best solution for the model. Considering only one variable (i.e., P) in the final objective function Eq.(7). To this end, we derived a straightforward and efficient solution. Specifically, Due to the basic roles of matrices, we have $\text{tr}(CE^\top PF) = \text{tr}(FCE^\top P)$. And we let $\mathcal{L}$ represent the value of Eq.(7), then we can obtain:

$$\mathcal{L} = \text{Tr}(FCE^\top P) - \mu_1 \|P - I\|_{\text{F}}^2 - \mu_2 \|P\|_{\text{F}}^2, \tag{8}$$

then we obtain the derivation of $\mathcal{L}$ with respect of P. The equation is shown below:

$$\frac{\partial \mathcal{L}}{\partial P} = (FCE^\top)^\top - 2\mu_1(P - I) - 2\mu_2 P, \tag{9}$$

To get the maximin value of Eq.(8), we assign the derivation of $\mathcal{L}$ to zero. The equation is $(FCE^\top)^\top - 2\mu_1(P - I) - 2\mu_2 P = 0$, and we eventually obtain the explicit solution of P which is shown below:

$$P = \frac{(FCE^\top)^\top}{2(\mu_1 + \mu_2)} + \frac{I}{1 + \frac{\mu_2}{\mu_1}}. \tag{10}$$

From Eq.(10), we can see that the solution of the final objective function Eq.(7) could be simple and straightforward. It is an explicit solution which only needs to be calculated once without further iteration. By the model design as well as the proposed solution, we could achieve effective and efficient domain adaptation in an open-set semantic segmentation scenario.

Experiments

In this section, we first give a brief introduction to the datasets and the evaluation metric used in our experiments. Then we present the implementation details of our proposed approach. Finally, we compare our results with various state-of-the-art models and conduct ablation studies to evaluate the effectiveness of different components of our proposed approach.

Experimental setup

Dataset For fair comparison, five classical and widely used semantic segmentation datasets are evaluated in our experiments. The brief introductions are listed below:

- **ADE20K-150** (Zhou et al. 2017) consists of 20,000 training images and 2,000 validation images and covers a wide range of scenes and contains annotations for a total of 150 classes.
- **ADE20K-847** (Zhou et al. 2017) shares the same images as ADE20K-150 but includes a larger set of annotated classes (847 classes), making it challenging for open-vocabulary semantic segmentation.
- **Pascal Context59** (Mottaghi et al. 2014) is a scene understanding dataset that comprises 5,000 training images, 5,000 validation images, and a total of 59 annotated classes.
- **Pascal Context-459** (Mottaghi et al. 2014) shares the same images as Pascal Context-59 but offers a larger set of annotated classes (459 classes). It is widely utilized for open-vocabulary semantic segmentation tasks.
- **Pascal VOC** (Everingham and Winn 2012) includes 20 classes of semantic segmentation annotations, with the training set comprising 1,464 images and the validation set containing 1,449 images.

Baseline methods Several state-of-the-art benchmarks are evaluated in our experiments. The brief introductions are listed below:

- **SimSeg†** (Xu et al. 2022a): A representative work of two-stage pixel-level semantic segmentation tasks, with the first stage obtaining mask proposals and the second stage performing open-vocabulary predictions based on CLIP model.

Table 1: Performance comparison with state-of-the-art methods.

Method	VL-Model	Training Set	ensemble.	ADE-847	PC-459	ADE-150	PC-59	VOC
SimSeg ECCV'22	CLIP ViT-B/16	COCO	no.	7.0	8.7	20.5	45.9	88.1
SimSeg† CVPR'23	CLIP ViT-B/16	COCO	yes.	6.9	9.7	21.1	52.2	92.3
OvSeg CVPR'23	CLIP ViT-B/16	COCO	yes.	7.1	11.0	24.8	52.7	92.5
SAN CVPR'23	CLIP ViT-B/16	COCO	no.	10.1	12.6	27.5	53.8	94.0
MADA (Ours)	CLIP ViT-B/16	COCO + 1% ADE-150	no.	**10.2**	**14.1**	**27.9**	**54.2**	**94.3**
MADA (Ours)	CLIP ViT-B/16	COCO + 1% VOC	no.	**10.2**	**14.2**	**28.0**	**54.2**	**94.4**
MaskCLIP ECCV'22	CLIP ViT-L/14	COCO	no.	8.2	10.0	23.7	45.9	-
SimSeg† CVPR'23	CLIP ViT-L/14	COCO	yes.	7.1	10.2	21.7	52.2	92.3
OvSeg CVPR'23	CLIP ViT-L/14	COCO	yes.	9.0	12.4	29.6	55.7	94.5
SAN CVPR'23	CLIP ViT-L/14	COCO	no.	12.4	15.7	31.9	57.7	94.6
MADA (Ours)	CLIP ViT-L/14	COCO + 1% ADE-150	no.	**12.8**	**16.6**	**31.9**	**57.8**	**95.6**
MADA (Ours)	CLIP ViT-L/14	COCO + 1% VOC	no.	**12.8**	**16.6**	**31.9**	**57.6**	**95.2**

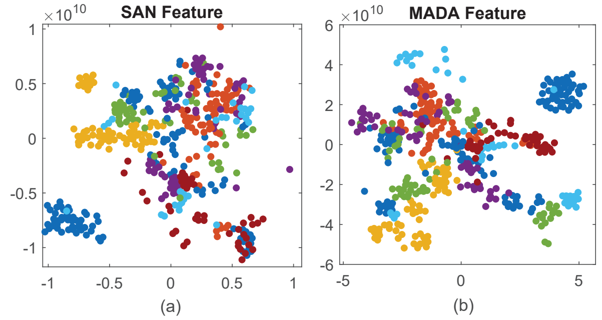

Figure 3: t-SNE visualization of the visual features. (a) denotes features extracted by SAN and (b) denotes features learned by our proposed MADA. Different colours represent different predicted categories. From the results, we can observe that our MADA achieve more distinguishing features between different compared with SAN.

- **MaskCLIP** (Ding, Wang, and Tu 2022): MaskCLIP represents a Transformer-based method that employs mask queries in conjunction with the ViT-based CLIP backbone for executing semantic segmentation and object instance segmentation.
- **OvSeg** (Liang et al. 2023): Due to the unsatisfactory mask proposal predictions of the pre-trained CLIP model, OvSeg performs mask prompt tuning with masked images and corresponding texts.
- **SAN** (Xu et al. 2023): SAN improves the heavy mask generator in MaskCLIP and OvSeg by decoupling mask recognition and mask prediction.

Evaluation In our experiments, we perform standard semantic segmentation metrics(Xu et al. 2023; Ghiasi et al. 2022; Xu et al. 2022b), i.e.the mean of class-wise intersection over union(mIoU) as metric to evaluate the performance of our model. mIoU is the average IoU over all classes. It takes the IoU value for each class, sums them up and then divides them by the number of classes. By taking the mean IoU over all classes, it ensures that all classes are considered equally, regardless of how often they appear in the dataset.

Implementation

We directly extract 512/768-dimensional mask proposal features and 512/768-dimensional word features from CLIP ViT-B/16 and CLIP ViT-L/14 SAN (Xu et al. 2023) models respectively to train. For each of the five datasets, a selection of different parameters of μ and the number of training images are chosen for training and further result comparison. The parameter sensitivity will be discussed in the following sections. All the training of our method is completed on a laptop with an Inter i7-8700 CPU and 32G memory. For a fair comparison, we test SAN models and our approach under the same environment: single RTX4090 GPU, i9-13900F CPU, 128G RAM, PyTorch 1.11.0, and CUDA 11.3.

Performance

Quantitative performance: Table 1 shows a quantitative comparison of our MADA with four state-of-the-art benchmarks. It can be seen that our MADA clearly surpasses all these methods, especially on PC-459 dataset under the same setting with an average of $+0.6\%$ mIoU for CLIP ViT-B/16 with only 100 training samples, which is only 1% from VOC training set. The results demonstrates the significant data usage efficiency of our MADA approach.

Feature Distribution Visualization: To further illustrate the feature distribution differences, t-SNE[1] is used for visualization and the result is shown in Figure 3, where different colors denotes different categories. From Figure 3, we observe that the feature distributions across different classes are more distinguishing compared with raw SAN features. This visualization explicitly reveals strong capacity of MADA for learning from target domains.

Parameter Sensitivity Analysis: We consider the training number and the first hyper-parameter μ_1 are crucial to determine MADA's performance. To this end, we analyze the parameter sensitivity and the results are illustrated in Figure 4, where we evaluated the training number from 10 to 200 on VOC dataset, and μ_1 from 20 to 2000. From Figure 4, we observed that the MADA could achieve high performance even with 50 samples which is only 0.5% of the VOC training set. It demonstrates the efficient data usage of MADA.

[1] https://lvdmaaten.github.io/tsne/

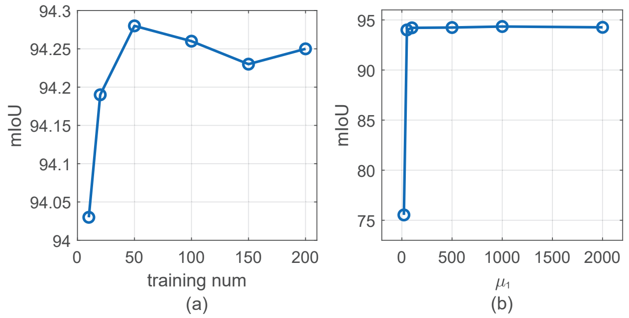

Figure 4: Parameter sensitivity analysis. (a) The performance of MADA as the training number increases from 10 to 200 on VOC dataset. (b) The performance of MADA as μ_1 changes from 20 to 2000. The plots demonstrates the data usage of MADA is efficient which achieves improvements with limited samples, and MADA is parameter-insensitive which is robust to different parameter/samples inputs.

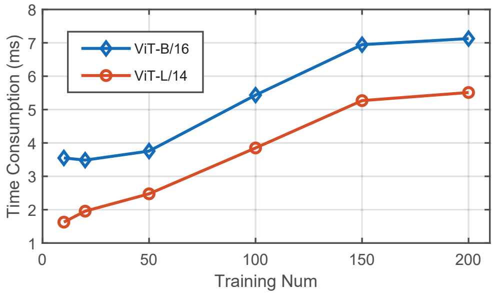

Figure 5: Time consumption of the optimization phrase of our MADA approach, based on both CLIP ViT-B/16 and CLIP ViT-L/14 respectively on VOC dataset. MADA optimization solution achieves considerable efficiency (millisecond-level) compared with other optimization methods.

Moreover, MADA is parameter-insensitive which have stable performance in a wide range of parameter μ_1 setup, it proves the robustness of MADA to different parameter/samples inputs.

Computational Efficiency Analysis: To evaluate the solution efficiency of MADA, we evaluate the time consumptions of solving Eq. (7) in different model/dimensional size and training numbers scenario, and the results are illustrated in Figure 5. Specifically, the both CLIP ViT-B/16 and CLIP ViT-L/14 backbones with different feature dimensions, as well as training numbers from 5 to 200 are tested. From Figure 5 we can see MADA is able to obtain results in less than 10 ms. This is due to our efficient solution and the parallel natural of matrix calculation in hardware.

Case Studies: Figure 6 illustrates the case studies of the segmentation results. Specifically, we use the same vocabulary and images from ADE20k-847 to test the model trained on COCO and 100 samples on ADE20k-150 dataset, and we can see that the "*floor*" is correctly classified by our MADA model but is misclassified as "*rug/carpet/carpe*" by SAN model. This case study further demonstrates the efficiency of MADA data usage.

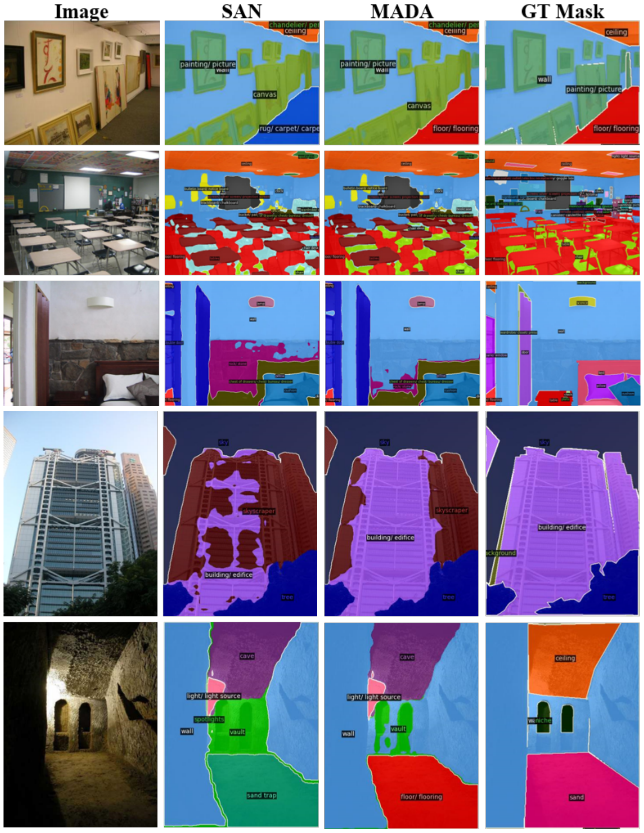

Figure 6: Oualitative examples demonstrating the effectiveness of MADA in improving mask proposal classification performance on ADE20k-847. The second column shows SAN inference results. We can see that MADA(third column) can improve semantic segmentation compared with the ground truth (fourth column). The colour setting is the same for all mask classes.

Conclusion

Our MADA framework addresses open-set semantic segmentation by efficiently adapting the model to the target domain using limited training samples. It effectively preserves the model's open-set capacity and performance. Our approach includes a Mask-Aware module to explore correlations between visual/mask space and feature space, a Structure Consistent module for stabilizing learning and maintaining generalization, and an efficient optimization solution. Extensive experiments verify the effectiveness of our approach.

Acknowledgments

The work was supported by China Southern Power Grid Digital Power Grid Research Institute Co., Ltd., and the project NO. is 210005KK52220019.

References

Brown, T.; Mann, B.; Ryder, N.; et al. 2020. Language models are few-shot learners. *Proceedings of NeurIPS*, 33: 1877–1901.

Bucci, S.; Loghmani, M. R.; and Tommasi, T. 2020. On the Effectiveness of Image Rotation for Open Set Domain Adaptation. In *Proceedings of ECCV*.

Bucher, M.; Vu, T.-H.; Cord, M.; and Pérez, P. 2019. Zero-shot semantic segmentation. *Proceedings of NeurIPS*, 32.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL*.

Ding, Z.; Wang, J.; and Tu, Z. 2022. Open-Vocabulary Universal Image Segmentation with MaskCLIP. *arXiv preprint arXiv:2208.08984*.

Everingham, M.; and Winn, J. 2012. The PASCAL visual object classes challenge 2012 (VOC2012) development kit. In *Pattern Analysis, Statistical Modelling and Computational Learning, Tech. Rep.*

Fu, C.; Chen, P.; Shen, Y.; Qin, Y.; Zhang, M.; Lin, X.; Qiu, Z.; Lin, W.; Yang, J.; Zheng, X.; et al. 2023. MME: A Comprehensive Evaluation Benchmark for Multimodal Large Language Models. *arXiv preprint arXiv:2306.13394*.

Geng, C.; Huang, S.-j.; and Chen, S. 2020. Recent advances in open set recognition: A survey. *IEEE Transactions on PAMI*, 43(10): 3614–3631.

Ghiasi, G.; Gu, X.; Cui, Y.; and Lin, T.-Y. 2022. Scaling Open-Vocabulary Image Segmentation with Image-Level Labels. In *Proceedings of ECCV*.

Jhoo, W.; and Heo, J.-P. 2021. Collaborative Learning With Disentangled Features for Zero-Shot Domain Adaptation. In *Proceedings of IEEE CVPR*.

Jing, T.; Liu, H.; and Ding, Z. 2021. Towards Novel Target Discovery Through Open-Set Domain Adaptation. In *Proceedings of ICCV*.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; Dollár, P.; and Girshick, R. 2023. Segment Anything. *arXiv:2304.02643*.

Lengyel, A.; Garg, S.; Milford, M.; and Gemert, J. 2021. Zero-Shot Day-Night Domain Adaptation with a Physics Prior. In *Proceedings of ICCV*.

Liang, F.; Wu, B.; Dai, X.; Li, K.; Zhao, Y.; Zhang, H.; Zhang, P.; Vajda, P.; and Marculescu, D. 2023. Open-vocabulary semantic segmentation with mask-adapted clip. In *Proceedings of IEEE CVPR*.

Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Li, C.; Yang, J.; Su, H.; Zhu, J.; et al. 2023. Grounding NIDO: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*.

Luo, Y.; Liu, P.; Guan, T.; Yu, J.; and Yang, Y. 2020. Adversarial Style Mining for One-Shot Unsupervised Domain Adaptation. In *Advances in Neural Information Processing Systems*.

Mottaghi, R.; Chen, X.; Liu, X.; Cho, N.-G.; Lee, S.-W.; Fidler, S.; Urtasun, R.; and Yuille, A. 2014. The Role of Context for Object Detection and Semantic Segmentation in the Wild. In *Proceedings of IEEE CVPR*.

Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; et al. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of ICML*, 8748–8763. PMLR.

Tzeng, E.; Hoffman, J.; Saenko, K.; and Darrell, T. 2017. Adversarial discriminative domain adaptation. In *Proceedings of IEEE CVPR*, 7167–7176.

Xian, Y.; Choudhury, S.; He, Y.; Schiele, B.; and Akata, Z. 2019. Semantic Projection Network for Zero- and Few-Label Semantic Segmentation. In *Proceedings of IEEE CVPR*.

Xu, M.; Zhang, Z.; Wei, F.; Hu, H.; and Bai, X. 2023. Side Adapter Network for Open-Vocabulary Semantic Segmentation. In *Proceedings of IEEE CVPR*.

Xu, M.; Zhang, Z.; Wei, F.; Lin, Y.; Cao, Y.; Han, H.; and Bai, X. 2022a. A Simple Baseline for Open Vocabulary Semantic Segmentation with Pre-trained Vision-language Model.

Xu, M.; Zhang, Z.; Wei, F.; Lin, Y.; Cao, Y.; Hu, H.; and Bai, X. 2022b. A simple baseline for open-vocabulary semantic segmentation with pre-trained vision-language model. In *Proceedings of ECCV*. Springer.

Yan, Z.; Wu, Y.; Li, G.; Qin, Y.; Han, X.; and Cui, S. 2022. Multi-level Consistency Learning for Semi-supervised Domain Adaptation. *arXiv preprint arXiv:2205.04066*.

Zhang, K.; Cai, R.; Wu, X.; Zhao, J.; and Qin, P. 2024. iBALR3D: imBalanced-Aware Long-Range 3D Semantic Segmentation.

Zhao, A.; Ding, M.; Lu, Z.; Xiang, T.; Niu, Y.; Guan, J.; and Wen, J.-R. 2021. Domain-Adaptive Few-Shot Learning. In *Proceedings of IEEE WACV*.

Zhou, B.; Zhao, H.; Puig, X.; Fidler, S.; Barriuso, A.; and Torralba, A. 2017. Scene Parsing through ADE20K Dataset. In *Proceedings of IEEE CVPR*.

Zou, X.; Dou, Z.-Y.; Yang, J.; Gan, Z.; Li, L.; Li, C.; Dai, X.; Behl, H.; Wang, J.; Yuan, L.; et al. 2023. Generalized decoding for pixel, image, and language. In *Proceedings of IEEE CVPR*, 15116–15127.