International Journal of Computer Vision
https://doi.org/10.1007/s11263-023-01863-1

Crots: Cross-Domain Teacher–Student Learning for Source-Free Domain Adaptive Semantic Segmentation

Xin Luo[1] · Wei Chen[1] · Zhengfa Liang[2] · Longqi Yang[3] · Siwei Wang[1] · Chen Li[1]

Received: 14 November 2022 / Accepted: 8 June 2023

Abstract

Source-free domain adaptation (SFDA) aims to transfer source knowledge to the target domain from pre-trained source models without accessing private source data. Existing SFDA methods typically adopt the self-training strategy employing the pre-trained source model to generate pseudo-labels for unlabeled target data. However, these methods are subject to strict limitations: (1) The discrepancy between source and target domains results in intense noise and unreliable pseudo-labels. Overfitting noisy pseudo-labeled target data will lead to drastic performance degradation. (2) Considering the class-imbalanced pseudo-labels, the target model is prone to forget the minority classes. Aiming at these two limitations, this study proposes a CROss domain Teacher–Student learning framework (namely CROTS) to achieve source-free domain adaptive semantic segmentation. Specifically, with pseudo-labels provided by the intra-domain teacher model, CROTS incorporates Spatial-Aware Data Mixing to generate diverse samples by randomly mixing different patches respecting to their spatial semantic layouts. Meanwhile, during inter-domain teacher–student learning, CROTS fosters Rare-Class Patches Mining strategy to mitigate the class imbalance phenomenon. To this end, the inter-domain teacher model helps exploit long-tailed rare classes and promote their contributions to student learning. Extensive experimental results have demonstrated that: (1) CROTS mitigates the overfitting issue and contributes to stable performance improvement, i.e., + 16.0% mIoU and + 16.5% mIoU for SFDA in GTA5→Cityscapes and SYNTHIA→Cityscapes, respectively; (2) CROTS improves task performance for long-tailed rare classes, alleviating the issue of class imbalance; (3) CROTS achieves superior performance comparing to other SFDA competitors; (4) CROTS can be applied under the black-box SFDA setting, even outperforming many white-box SFDA methods. Our codes will be publicly available at https://github.com/luoxin13/CROTS.

Keywords Domain adaptation · Transfer learning · Semantic segmentation

1 Introduction

Semantic segmentation (Asgari Taghanaki et al., 2021; Chen et al., 2017; Kamann & Rother, 2021; Hu et al., 2022; Yu et al., 2022) assigns semantic labels for each pixel in the given image, which plays a vital role in autonomous driving, scene understanding, etc. Recent advances in semantic segmentations largely attribute to deep learning methods. Despite the promising performance, these methods are heavily dependent on a large quantity of labeled training data, which is expensive and time-consuming to collect.

Communicated by Ming-Hsuan Yang.

Xin Luo and Wei Chen have contributed equally to this work.

✉ Wei Chen
chenwei@nudt.edu.cn

Xin Luo
luoxin13@nudt.edu.cn

Zhengfa Liang
liangzhengfa10@nudt.edu.cn

Longqi Yang
yanglongqi19@nudt.edu.cn

Siwei Wang
wangsiwei13@nudt.edu.cn

Chen Li
lichen14@nudt.edu.cn

[1] College of Computers, National University of Defense Technology, Changsha, China

[2] National Key Laboratory of Science and Technology on Blind Signal Processing, Chengdu, China

[3] Defense Innovation Institute, Beijing, China

Published online: 18 August 2023

Considering the pricey cost of collecting labeled training data, it has long been desired to train the task model with label-rich source domains and adapt it to a label-scarce target domain. To this end, unsupervised domain adaptation (UDA) methods (Tsai et al., 2018; Zou et al., 2018, 2019; Saito et al., 2018; Vu et al., 2019; Lee et al., 2019; Li et al., 2019; Chang et al., 2019; Li et al., 2021; Zhao et al., 2021; Yu et al., 2021) have been proposed to bridge the gap between the source and target domains. Typical UDA methods benefit from statistical distribution alignment or adversarial learning, which maps cross-domain data into a domain-invariant feature space. Learning domain-invariant feature representation generally requires the coexistence of source and target data. Nevertheless, access to source data cannot always be guaranteed due to privacy, storage, transmission, and other issues, making UDA methods unworkable. To address these issues, source-free domain adaptation (SFDA) methods (Ye et al., 2021; You et al., 2021; S & Fleuret, 2021; Huang et al., 2021; Nath Kundu et al., 2020; Kundu et al., 2021; Wang et al., 2022) relax the reliance on source data by adapting from pre-trained source models. Mainstream SFDA methods are based on self-training paradigms. They utilize pre-trained source models to generate pseudo-labels for unlabeled target data. The pseudo-labeled target data are exploited to re-train the task model. Then, the iterative updating between pseudo-labels and the task model helps promote the adaptation.

Despite the promising improvements, existing SFDA methods still suffer from two limitations: (1) Pseudo-labeling generally assumes that unlabeled data are sampled from a similar distribution with labeled data. However, in practice, there usually exists a large discrepancy of distribution between source and target data domains, which leads to severe noise in target pseudo-labels and subsequently contaminates self-training performance. (2) Semantic segmentation naturally bears long-tailed imbalanced class distributions. For instance, in GTA5 (Richter et al., 2016), a synthetic urban-scene dataset for semantic segmentation, the *road* class may occupy over *30%* of the whole pixels in an image, while the *bike* class only accounts for *0.01%*. The discrepancy between different domains exacerbates such imbalance. Under the SFDA setting, without accurate supervision from source data, learning with imbalanced pseudo-labels results in the model's failure to perform well for the minority classes.

Aiming at these two limitations, we design a CROss domain Teacher–Student (namely CROTS) learning framework for source-free domain adaptive semantic segmentation. Specifically, CROTS conducts teacher–student learning in the target domain, where the temporal ensemble of the student models is used as the intra-domain teacher model to predict pseudo-labels for unlabeled target data. During learning, pseudo-labeled target images are separated into small patches. CROTS generates novel training samples by mixing these patches that belong to different images. It should be noted that the mixing is spatial-aware, which maintains the spatial layout of image patches and thus regularizes spatial prior. Spatial-aware data mixing benefits the task model in two aspects: (1) The generated samples integrate statistics of different images. Since data statistics are related to the style of data domains, with generated samples, CROTS expands the style diversity of the training data, which helps alleviate the overfitting issue. (2) Patch-wise image mixing changes the background of the original image patches, which promotes contextual consistency for the task model, i.e., forcing the task model to produce consistent predictions for a given image patch, regardless of its contextual background patches.

Meanwhile, considering the severe class imbalance, intra-domain teacher–student learning in the target domain is dominated by the majority classes, suffering from significant performance degradation for the minority classes. Therefore, as the learning proceeds, for those minority classes, the pre-trained source model will perform better than the intra-domain teacher model. Based on this observation, CROTS employs the pre-trained source model as the inter-domain teacher, which provides inter-domain pseudo-labels for unlabeled target data and helps mine patches that contain long-tailed rare classes. During learning, these patches augment the contribution of those minority classes, which mitigates the issue of class imbalance.

Overall, the proposed CROTS framework incorporates intra-domain data mixing and inter-domain patch mining to achieve cross-domain teacher–student learning, which addresses the two limitations in the current SFDA methods, i.e., the overfitting issue and the class imbalance issue. Besides, since CROTS is only dependent on the outputs of teacher models, it can be applied to adapt black-box source models. Under the black-box SFDA setting, the target model cannot use the pre-trained source model as initialization, and access to the weights of the pre-trained model is prohibited. Instead, the only accessible resource from the source vendor is an interface that provides pseudo-labels for given input data.

Our contributions are summarized as follows:

- A CROss-domain Teacher–Student (CROTS) learning framework is proposed for source-free domain adaptive semantic segmentation.
- A data augmentation strategy is fostered to avoid the overfitting issue in SFDA, which enriches the diversity of unlabeled target training data via mixing various training samples while respecting their spatial semantic layouts.
- A training regularization is designed to mitigate the issue of class imbalance, which exploits the inter-domain teacher model for patch samples of rare long-tailed classes to augment their contributions for domain adaptation.

- Experimental results confirm that CROTS outperforms many leading SFDA methods. Additional experiments are performed to explore the extension of CROTS under the black-box SFDA setting, which verifies CROTS's capability of adapting black-box source models.

The rest of this paper is organized as follows. Section 2 outlines the related work of domain adaptive semantic segmentation. Section 3 presents the proposed CROTS framework. Further, we also provide a detailed description of Spatial-Aware Data Mixing and Rare-Class Patches Mining. Section 4 shows the experiment results with evaluation and ablation. Section 5 concludes the paper.

2 Related Work

Numerous attempts have been made to achieve domain adaptive semantic segmentation. In this section, we briefly review literature closely related to our work, which covers typical methods in domain adaptation and data augmentation.

2.1 Unsupervised Domain Adaptation

Unsupervised Domain Adaptation (UDA) utilizes labeled source data and unlabeled target data to perform cross-domain tasks (Ganin et al., 2016; Du et al., 2021; Luo et al., 2019; Lu et al., 2020; Lee et al., 2019; Li et al., 2021; Zhao et al., 2021). Early UDA studies for semantic segmentation benefit from domain adversarial learning (Tsai et al., 2018; Saito et al., 2018). With the adversary between the feature extractor and the domain discriminator, images from different domains are mapped into a shared latent space, contributing to domain-invariant task learning and promoting the alignment of different domains (Vu et al., 2019; Chang et al., 2019; Yang et al., 2020; Lu et al., 2020; Du et al., 2021).

Another line of studies tackles the UDA problem within the semi-supervised learning framework, where pseudo-labeling is introduced to simulate supervised learning in the unlabeled target domain. With considerable techniques for improving the quality of pseudo-labels (Zou et al., 2018, 2019; Ganin et al., 2016; Zhang et al., 2021; Zheng & Yang, 2021), pseudo-labeling has significantly boosted the performance of unsupervised domain adaptive semantic segmentation.

However, typical UDA methods require the co-occurrence of source and target data, which is not always applicable due to privacy, storage, transmission, and other concerns.

Besides, some attempts have been made to alleviate the issue of class-imbalance by incorporating cross-domain mixed sampling (Tranheden et al., 2021), class-balanced pseudo labeling (Li et al., 2022), rare-class sampling (Hoyer et al., 2022), etc. Despite the contribution of these methods to alleviating the issue of class imbalance, they are conducted with the help of source annotation, which is not applicable when source data is not accessible.

2.2 Source-Free Domain Adaptation

Source-Free Domain Adaptation (SFDA) achieves the same goal as UDA while avoiding access to source data (Ahmed et al., 2021; Nath Kundu et al., 2020; Yang et al., 2021; Kurmi et al., 2021; Liang et al., 2020; Li et al., 2020). Similar methods have also been proposed for the task of semantic segmentation. Typical SFDA methods for semantic segmentation (Ye et al., 2021; You et al., 2021; Huang et al., 2021; S & Fleuret, 2021; Liu et al., 2021; Wang et al., 2021) apply self-training to promote the target adaptation. Some methods update the target model by minimizing the prediction entropy (Wang et al., 2021), while most other methods benefit from pseudo-labeling. Without the help of the source domain, pseudo-labeling often suffers from the overfitting issue and the bias toward the majority classes.

Black-Box Source-Free Domain Adaptation (B2SFDA) is a specific example of SFDA, which prohibits direct access to the source model weights. Under the B2SFDA setting, the only accessible resource is an API service, where the client feeds the interface with target samples, and the source vendor returns the corresponding predictions. Some studies achieve black-box adaptation for image classification via knowledge distillation(Zhang et al., 2021; Liang et al., 2022). However, B2SFDA is less explored for the task of semantic segmentation, which needs pixel-level annotations and is much harder. In this study, we extend our proposed method under the black-box setting, which provides a general solution to black-box source-free domain adaptation in semantic segmentation.

2.3 Data Augmentation

Data augmentation has been a common practice to increase the diversity of training data and regularize the learning process to avoid overfitting. In the context of semantic segmentation, many techniques have been employed to augment images(French et al., 2020a). Image-level mixing, i.e., Mixup (Zhang et al., 2018), mixes two images via pixel-wise linear combination. CutMix (Yun et al., 2019) is proposed to combine two images in a patch-wise manner. To enhance the task relevance, ClassMix (Olsson et al., 2021) is proposed to improve CutMix, which leverages model predictions to exploit the semantic boundaries of mixing patches. CowMix (French et al., 2020b) mixes different images in a randomized manner. These methods mix images to synthesize novel training samples. However, most of these methods only mix two images and neglect the spatial contexts of semantic patches (e.g., pasting a *car* patch onto a *sky* background). Besides, state-of-the-art mixing strategies are

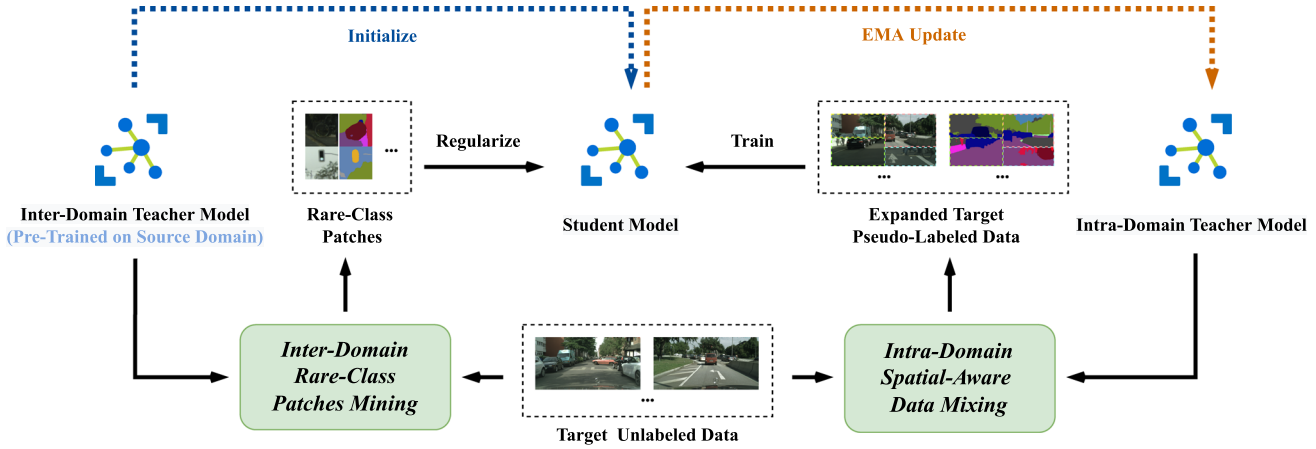

Fig. 1 Overview of the proposed CROTS. CROTS conducts teacher–student learning in the target domain. Specifically, the intra-domain teacher model is the temporal mean of the student models, which generates pseudo-labels for unlabeled target data. CROTS introduces a novel data augmentation strategy named Spatial-Aware Data Mixing to enhance data diversity. Meanwhile, the pre-trained source model is utilized as the inter-domain teacher model. CROTS designs the Rare-Class Patches Mining mechanism to exploit image patches of long-tailed classes

dedicated to semi-supervised semantic segmentation, which generally assumes that unlabeled data are subject to similar distribution with labeled data. For domain adaptive semantic segmentation, these methods may suffer from limited performance gain. In this study, we propose to mix more images while maintaining the spatial layouts of mixing patches, which works well for domain adaptive semantic segmentation.

2.4 Teacher–Student Learning

Teacher–Student learning is first introduced in knowledge distillation (Gou et al., 2021), where the student model imitates the behavior of the teacher model. Tarvainen and Valpola (2017) introduced the Mean-Teacher model to utilize the exponential moving averaged student model as the teacher model, which serves as the temporal ensembling of historical student models. Liu et al. (2022) introduce a perturbed auxiliary teacher model to improve the Mean-Teacher model, which outperforms state-of-the-art semi-supervised methods for semantic segmentation. In this study, the Mean-Teacher model is introduced as our intra-domain teacher model. Besides, we employ the pre-trained source model as the inter-domain teacher model. The cooperation of intra-domain and inter-domain teachers helps overcome the two limitations of SFDA.

3 Cross-Domain Teacher–Student Learning

This section will cover the details of the proposed Cross-Domain Teacher-Student (CROTS) Learning Framework, which is designed for source-free domain adaptive semantic segmentation. Considering the practical application scenes, we also extend CROTS for the black-box SFDA setting.

First, we give an overview of the proposed CROTS framework in Sect. 3.1. Then, we introduce an intra-domain teacher model to generate pseudo-labels and propose a novel data augmentation strategy to expand the scale of training data, which is demonstrated in Sect. 3.2. Next, we design a patch mining mechanism to regularize the task model for class balancing in Sect. 3.3. Finally, we describe the overall training objective of CROTS and extend it for the black-box setting, which will be covered in Sects. 3.4 and 3.5, respectively.

3.1 Overview

During adaptation, the source vendor provides a pre-trained source model ($\mathcal{M}_s$, which can be a black box). The target client has access to N_t unlabeled target images ($\mathcal{D}_t = \{x_t^{(i)}\}_{i=1}^{N_t}$). Based on $\mathcal{M}_s$ and $\mathcal{D}_t$, the target client aims to learn the task model $\mathcal{M}_t$, which can be used to generate semantic segmentation results for target data.

As shown in Fig. 1, the proposed CROTS conducts teacher–student learning in the target domain, where the temporal ensemble of the student model is used as the intra-domain teacher model to predict pseudo-labels for unlabeled target data. Meanwhile, CROTS employs the pre-trained source model as the inter-domain teacher, which provides inter-domain pseudo-labels for unlabeled target data and helps mine patches that contain long-tailed rare classes.

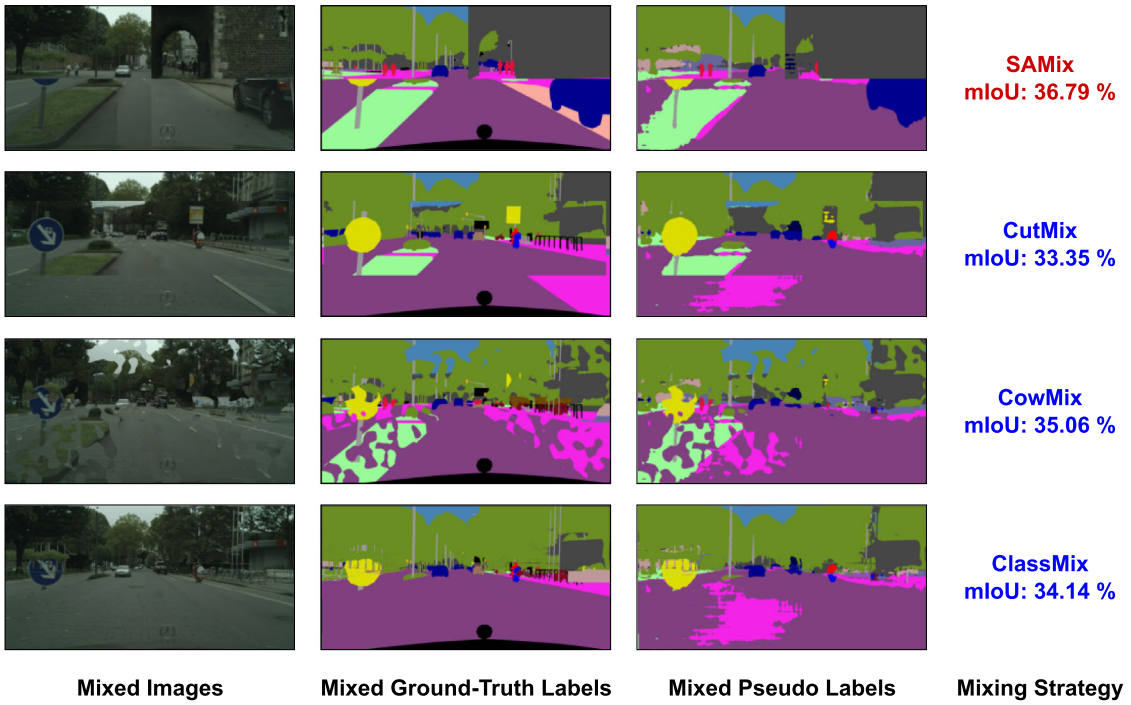

Fig. 2 Comparison among different mixing strategies. Each row gives the mixed images, the mixed ground-truth labels, and the mixed pseudo labels of a mixing strategy. Under the SFDA setting, considering the severe noise in pseudo labels, The proposed Spatial-Aware Data Mixing aims to increase the data diversity while mitigating the noise in pseudo labels

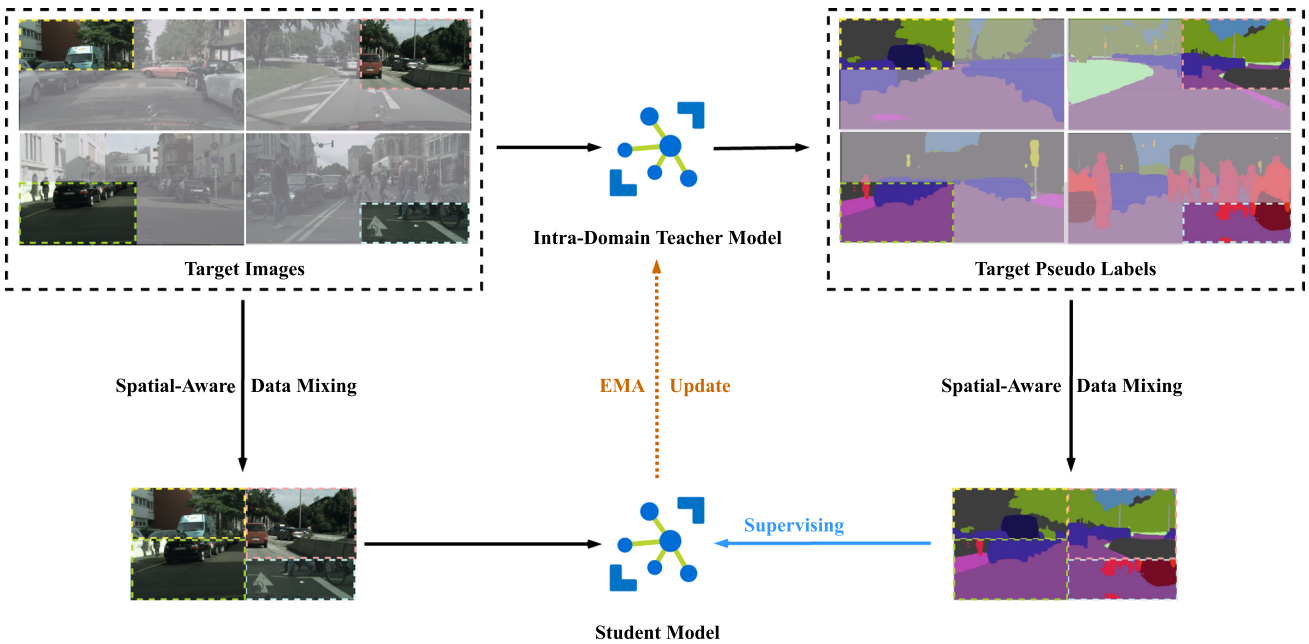

Fig. 3 Illustration of the *Spatial-Aware Data Mixing* module. The dashed colored rectangles denote the mixing masks sampled from a *Beta* distribution. We utilize these masks to mix several images. Then, we apply the same mixing strategy to their pseudo-labels. The proposed *Spatial-Aware Data Mixing* module effectively enriches the diversity of training data without significantly changing the spatial prior

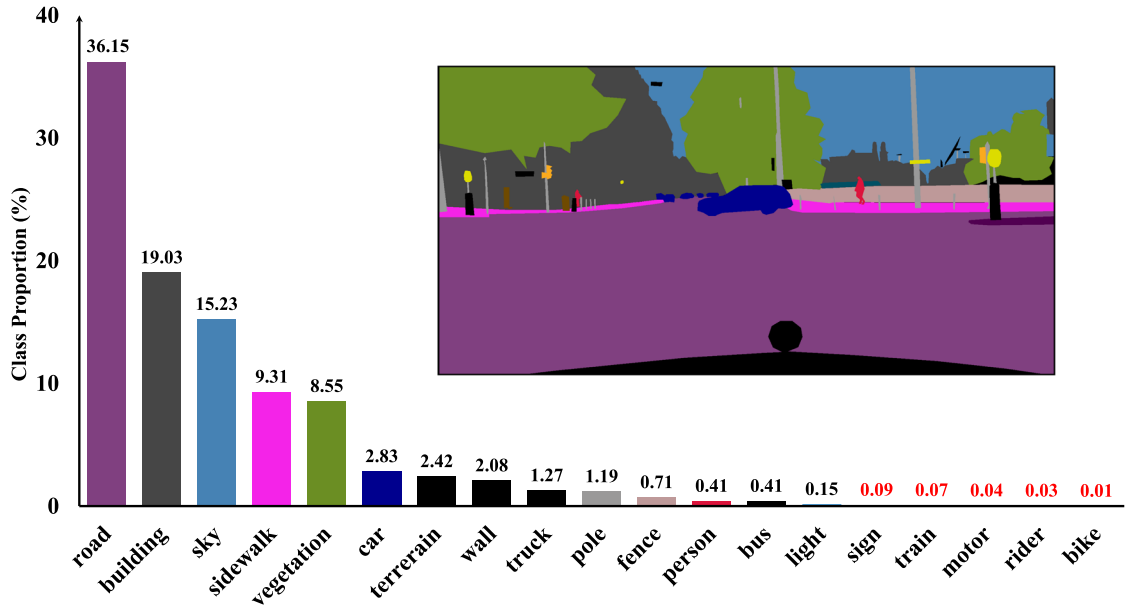

Fig. 4 Illustration of long-tailed distribution over different semantic classes. The statistics in the bar chart are calculated from the ground-truth labels in the GTA5 (Richter et al., 2016) dataset (a synthetic dataset for self-driving semantic segmentation). Proportion statistics less than 0.1% in the dataset are marked in red. Their corresponding classes are recognized as long-tailed rare classes. The image in the figure is an example of ground-truth segmentation labels, which illustrates the long-tailed class distribution at the image level (Color figure online)

3.2 Intra-Domain Learning

Under the source-free setting, the adaptation process is conducted only within the target domain. The intra-domain learning in the target domain is prone to overfit the noisy pseudo labels, leading to degraded performance.

In this section, aiming at the challenges in SFDA semantic segmentation, we introduce a novel mixing strategy for data augmentation, which enhances the performance of intra-domain learning.

3.2.1 Motivation

Data augmentation has been widely adopted to increase the diversity of training data and to avoid the overfitting issue. Among these data augmentation methods, mixing strategies have proved to be effective in generating novel samples.

existing mixing strategies are generally designed for the semi-supervised learning problem. For instance, ClassMix (Olsson et al., 2021) generates a new data-label pair by employing the predicted class-wise masks to mix two images, which achieves promising performance in semi-supervised semantic segmentation. Figure 2 compares the generated images and pseudo-labels of applying different mixing strategies in SFDA. From the figure, it can be observed that the pseudo-labeled class masks in SFDA are far from accurate, limiting the performance of previous mixing methods.

With regard to these issues, a novel data mixing strategy needs to tackle two problems before it can be well applied in SFDA semantic segmentation:

- *The Limited Diversity.* Typical mixing methods combine two images to produce a novel sample. To further boost sample diversity, the number of mixing samples needs to be increased.
- *The Ignored Spatial Prior.* In CutMix and Jigsaw (Bochkovskiy et al., 2020), the mixing patches are randomly re-scaled and then pasted onto random positions, which might break the spatial prior of semantic segmentation. For example, it is irrational to paste a *sky* patch onto a *road* patch. With this concern, the spatial prior must be taken into consideration, so that the spatial layouts of generated samples are similar to realistic samples.

3.2.2 Spatial-Aware Data Mixing

Aiming at the pitfalls of applying previous mixing strategies for SFDA, we design the Spatial-Aware Data Mixing module to augment target images.

Specifically, the proposed Spatial-Aware Data Mixing extends the CutMix method in two aspects: (1) We increase the number of mixed samples from 2 to N, which increases the diversity of target data. (2) We regularize the spatial prior of semantic classes by mixing the patches without changing their spatial layouts.

Table 1 Performance comparison of *CPSL* under different settings

Method	road	sidewk.	build	wall	fence	pole	light	sign	vege	terr	sky	person	rider	car	truck	bus	train	motor	bike	**mIoU**
Source Only	75.8	16.8	77.2	12.5	21.0	25.5	30.1	20.1	81.3	24.6	70.3	53.8	26.4	49.9	17.2	25.9	6.5	25.3	36.0	36.6
CPSL (UDA)	92.3	59.9	84.9	45.7	29.7	52.8	61.5	59.5	87.9	41.5	85.0	73.0	35.5	90.4	48.7	73.9	26.3	53.8	53.9	60.8
CPSL (SFDA)	87.7	48.7	84.2	37.0	35.6	46.6	49.0	53.6	87.9	45.1	84.9	60.8	0.0	88.4	25.3	48.4	0.0	0.0	44.7	*48.8*
CROTS (SFDA)	87.6	42.6	85.7	22.5	38.9	38.3	50.3	56.2	88.0	38.8	90.7	68.3	38.7	83.5	20.7	40.3	11.7	42.7	60.6	**53.0**

The results of *Source Only* are borrowed from (Vu et al., 2019). The results of *CPSL (UDA)* are from (Li et al., 2022). The results of *CPSL (SFDA)* are based on our re-implementation. UDA methods suffers from significant performance degradation under the SFDA setting

Figure 3 illustrates the design of Spatial-Aware Data Mixing. Let x_t be the target image to be augmented. First, we randomly sample the position masks to split the target images into N patches, where any two masks are disjoint, and the union of all masks corresponds to the whole image. Then, we randomly sample $N-1$ images from the target domain. According to the position masks, the sampled $N-1$ images are pasted onto x_t. Meanwhile, we apply the same operation to produce pseudo-labels for mixed images. The proposed Spatial-Aware Data Mixing can be formulated as:

$$\begin{aligned} x_t^{mix} &= (1 - \bigcup_{i=1}^{N-1} m_i) \odot x_t + \sum_{i=1}^{N-1} m_i \odot x_t^{(i)}; \\ \tilde{y}_t^{mix} &= (1 - \bigcup_{i=1}^{N-1} m_i) \odot \tilde{y}_t + \sum_{i=1}^{N-1} m_i \odot \tilde{y}_t^{(i)}, \end{aligned} \tag{1}$$

where m_i denotes the mask for mixing the i-th images, and $\tilde{y} = \text{argmax}_c \overline{\mathcal{M}}_t(x)$ is the one-hot label predicted by the mean teacher model $\overline{\mathcal{M}}_t$. To obtain random spatial masks, we sample a random height and a random width from a *Beta* distribution, which separates x_t into four parts. Then, patches from the same location in the other three images are copied and pasted onto the target image (as illustrated in Fig. 3).

3.3 Inter-Domain Learning

Class imbalance is a long-standing issue in semantic segmentation. Semantic classes are generally long-tail distributed. Figure 4 gives the proportion of each classes in the GTA5 dataset (Richter et al., 2016). As shown in the figure, the head classes such as *road* occupy the majority of all pixels in the dataset. As a contrast, the tail classes such as *bike* only account for about *0.01%* of the whole dataset.

In this section, to mitigate forgetting long-tailed rare classes, we employ the pre-trained source model as an inter-domain teacher model, which increases the contribution of rare classes. During inter-domain learning, we introduce a novel class-balancing strategy to mine rare-class samples from a pre-trained model. These samples are utilized to regularize class balance.

3.3.1 Motivation

With imbalanced class distributions, rare classes provide limited supervision for training the task model. Besides, with the sharp discrepancy between source and target domains, samples of long-tailed rare classes in the source domain are almost impossible to represent samples of the same class in the target domain. This poses a strict challenge for domain adaptive semantic segmentation, where the adapted task model suffers from a performance drop for those long-

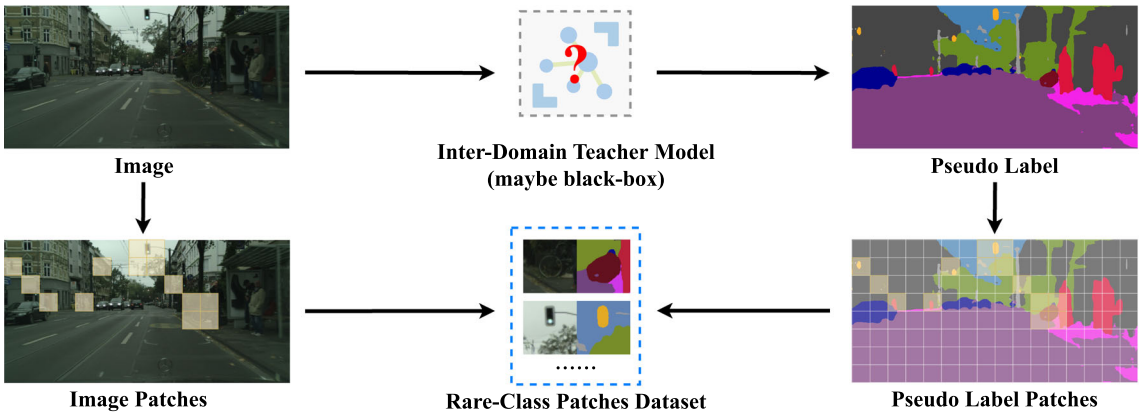

Fig. 5 Illustration of rare-class patches mining. Given an unlabeled target image, we obtain its pseudo-label from the pre-trained source model. Then, the image and its corresponding pseudo-label are split into small patches. Next, patches containing pixels of rare classes are recognized as rare-class patches. Finally, those rare-class patches are stored to regularize the training process in SFDA

tailed classes. For instance, after adaptation, the IoU metric of the *train* class decreased to 0 (as shown in Table 4). Regarding this issue, some training strategies are required to increase the contribution of long-tailed rare classes and to achieve class balancing.

Despite the effort of UDA methods to mitigate class imbalance, these methods may fail to work in the source-free setting. As listed in Table 1, when source data is absent, the performance of CPSL (Li et al., 2022) decreases significantly, with mIoU from 60.8% to 48.8%. This indicates that UDA methods cannot be directly adopted in the SFDA setting. Therefore, a novel method needs to be proposed to tackle the imbalanced class distributions under the SFDA setting.

3.3.2 Rare-Class Patches Mining

With the imbalanced class distribution and the noisy pseudo labels, the adapted task model in SFDA suffers from degraded performance for long-tailed rare classes. Under this circumstance, the pre-trained source model can perform better to segment rare-class pixels. Therefore, the pre-trained source model can provide useful information for long-tailed rare classes.

Considering this, we employ the pre-trained source model as the inter-domain teacher model. Based on this teacher model, we propose a class-balancing mechanism to assist with target adaptation, namely Rare-Class Patches Mining.

Rare-Class Patches Mining utilizes the inter-domain teacher model to generate inter-domain pseudo-labels. With these pseudo-labeled data, patches containing long-tailed classes are found as rare-class patches. Figure 5 illustrates the operation of Rare-Class Patches Mining, which generates pseudo-labeled image patches for long-tailed classes and marks them as rare-class patches.

Specifically, we first rank all the classes according to the number of predicted pixels for each class. For the N_{rare} ($N_{rare} = 8$ in our experiments) classes with the least number of pixels, we mark them as rare classes. Then, we split each target image into patches of size $R_h \times R_w$ (set to be 64×64 in all experiments). Among those patches, if they contain pixels of rare classes, we recognize them as rare-class patches and store them for regularizing teacher–student learning.

During learning, CROTS employs rare-class patches as an additional training dataset. Concretely, the sampled patches are pasted to a standard image to guarantee reasonable spatial resolution. After the forward pass, the loss calculation is only applied to the regions where rare-class patches are pasted. Rare-class patches force rare classes to make more contributions to loss calculation, mitigating the class imbalance.

3.4 Overall Training Objective

During teacher–student learning, CROTS utilizes the diversified intra-domain pseudo-labeled data to train the task model. Meanwhile, rare-class patches from inter-domain pseudo-labels contribute to an auxiliary loss. Overall, the learning objective of CROTS is formulated as follows:

$$\begin{aligned}\mathcal{L}_{crots} = &\frac{1}{N_t}\sum_{i=1}^{N_t}\mathcal{L}_{ce}(x^i_{mix}, y^i_{mix}; \mathcal{M}_t) \\ &+\frac{1}{N_r}\sum_{i=1}^{N_h}\mathcal{L}_{ce}(x^i_{rare}, y^i_{rare}; \mathcal{M}_t),\end{aligned} \tag{2}$$

where the pair (x^i_{mix}, y^i_{mix}) denotes a novel sample pair generated from Spatial-Aware Data Mixing, the pair (x^i_{rare}, y^i_{rare})

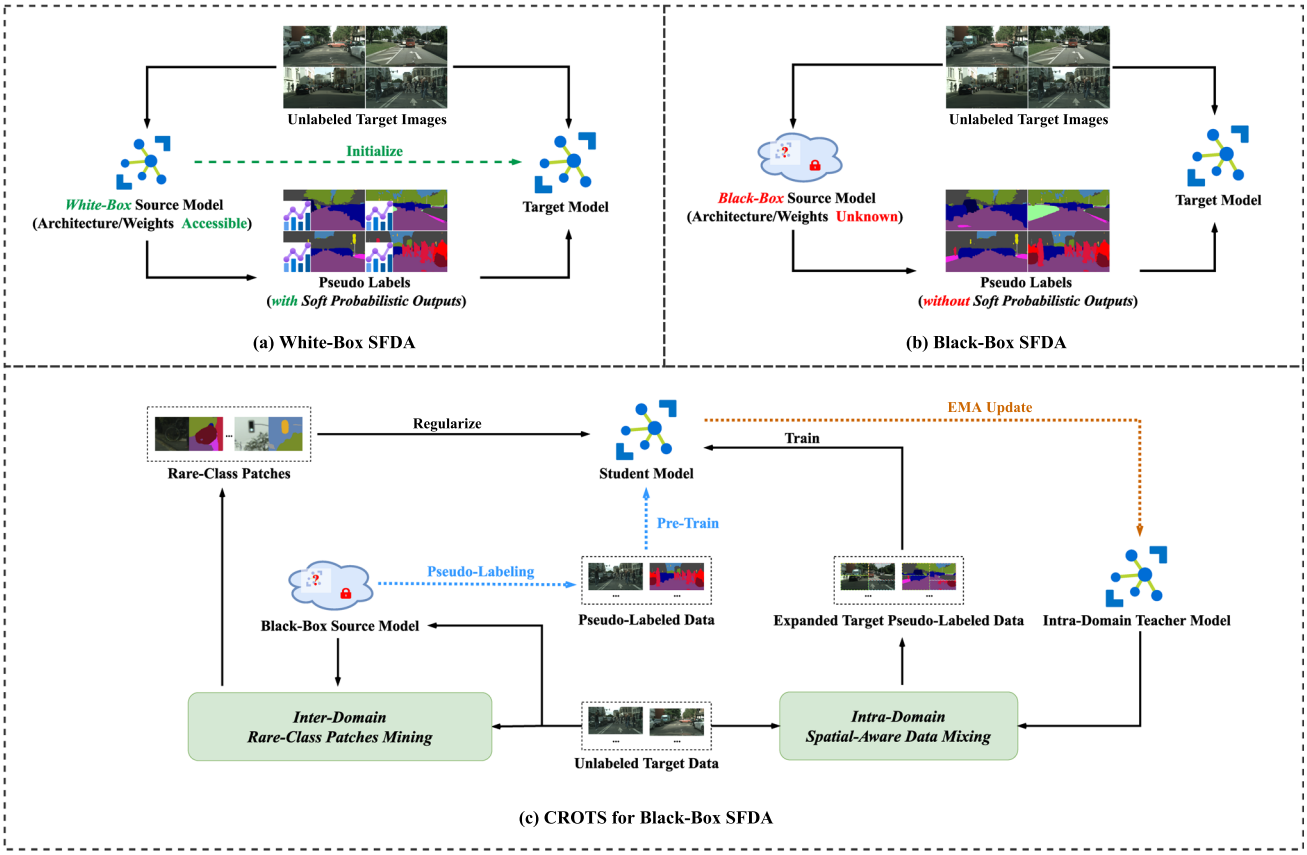

Fig. 6 Extending CROTS for black-box SFDA. **a** White-Box SFDA initializes the target model with the weights of the source model and has full access to the source models. **b** In contrast, Black-Box SFDA only utilizes the source model as an API service to get naive pseudo labels (*without* soft probabilistic outputs), without access to either details or weights of the source model. **c** CROTS can be extended for Black-Box SFDA

is a rare-class patch and its pseudo-label, and N_r denotes the total number of rare-class patches.

3.5 Extension for Black-Box Setting

In this section, we introduce how CROTS can be applied for Black-Box SFDA.

Under the Black-Box SFDA setting, to save the cost of transmission and storage, the client uploads target images to the source vendor, which only provides naive pseudo-labels (*without the soft probabilistic outputs*) for target adaptation.

Since Rare-Class Patches Mining and Spatial-Aware Data Mixing only depend on the image and its inter-domain pseudo-labels, they can be used for the Black-Box setting without modification. However, to achieve effective teacher–student learning, CROTS relies on an initialized student model for better performance, which is not available under the Black-Box setting. To this end, this study extends the proposed CROST by casting the Black-Box SFDA problem as two-stage knowledge distillation. The overview of the extended CROTS is depicted in Fig. 6. In the first stage, cross-domain distillation extracts pseudo-labeled target data from the black-box source model. These pseudo-labeled target data pre-train the student model to imitate the black-box source model. After this, the pre-trained student model can be used as initialization and enables CROST to work in the same manner as mentioned earlier.

4 Experiments

In this section, we conduct experiments to validate the effectiveness of the proposed *CROTS*. Overall, this section aims to answer the following questions:

- *RQ1:* Does *CROTS* manage to tackle the SFDA problem? How does it perform compared with other SFDA or UDA methods?
- *RQ2:* Does *Spatial-Aware Data Mixing* avoid overfitting and promote performance?

- *RQ3:* Does *Rare-Class Patches Mining* increase the contribution of long-tailed rare classes and alleviate the issue of class imbalance?
- *RQ4:* How do hyper-parameters affect *CROTS*? What are the optimal hyper-parameters? Is *CROTS* sensitive to different source models?
- *RQ5:* How does the proposed *CROTS* perform to adapt black-box pre-trained models?

To answer the above questions, first, we introduce details of the benchmark as well as the evaluation metrics. Then, we cover the implementation details. Next, we report the performance of the proposed method and compare it against state-of-the-art SFDA methods. Finally, we also conduct ablation studies to give an insightful investigation of the proposed CROTS.

4.1 Setup

4.1.1 Datasets

Cityscapes (Cordts et al., 2016) is a well-adopted benchmark for urban-scene semantic segmentation. It collects urban-scene images from 50 European cities, consisting of a training split of 2975 samples, a testing split of 888 samples, and a validation split of 500 samples. In our experiments, we employ it as the target dataset. We only have access to unlabeled images in the training split during training. **GTA5** (Richter et al., 2016) is a synthetic dataset generated from computer gaming engines, which contains 24,966 labeled images. Similarly, **SYNTHIA-RAND-CITY** (Ros et al., 2016) is also a synthetic dataset generated from computer programs, consisting of 9600 labeled images. These two synthetic datasets have compatible annotations with the Cityscapes dataset. In previous studies, they have been widely employed as the source datasets.

4.1.2 Evaluation Metric

Following previous studies in semantic segmentation, we evaluate and report the performance of the Cityscapes validation split. Specifically, we report both the class-wise *intersection-over-union (IoU)* metrics and the *mean IoU (mIoU)* over all the classes. mIoU is calculated as:

$$\text{mIoU} = \frac{1}{C}\sum_{i=1}^{C}\frac{p_{ii}}{(\sum_{j=1}^{C} p_{ij}) + (\sum_{j=1}^{C} p_{ji}) - p_{ii}}, \tag{3}$$

where p_{ij} denotes the number of pixels that belong to the *i-th* class and are predicted as the *j-th* class.

4.1.3 Implementation Details

Following most UDA and SFDA studies for semantic segmentation, we employ the *DeepLabV2* framework (Chen et al., 2017) as the segmentation network, with the *ResNet101* (He et al., 2016) pre-trained on ImageNet (Russakovsky et al., 2015) as the backbone. The whole solution is implemented on the Pytorch platform (Paszke et al., 2019) and runs on a single NVIDIA RTX 6000 GPU with 24 GB memory. The whole network is optimized via the stochastic gradient descent (SGD) optimizer (Summa et al., 2010), with a momentum of 0.9 and a weight decay factor of 5×10^{-4}. The initial learning rate is 5×10^{-5} and is scheduled using the Cosine scheduler (Loshchilov & Hutter, 2017). During training, we resize target images to 1024×512. The batch size is set to 4.

4.1.4 Baseline Methods

(a) *Source Only* model is the model without adaptation. Instead of training a source model from scratch, we directly employ the pre-trained source models of AdaptSegNet (Tsai et al., 2018) and GtA (Kundu et al., 2021), where the former is the basic source model and the latter is a generalized source model (which applies domain generalization techniques to improve the performance).

(b) *UDA methods* are dependent on the co-existence of source and target data. During the comparison, we choose typical UDA methods as counterparts: AdvEnt (Vu et al., 2019) is based on adversarial learning), while CRST (Zou et al., 2019), DAST (Yu et al., 2021), and MRNet+(Zheng & Yang, 2021) are based on self-training. In addition to these typical methods, we also list state-of-the-art UDA methods, e.g., CPSL (Li et al., 2022), DAFormer (Hoyer et al., 2022), etc., which demonstrate the developing trend in the whole domain adaptation community.

(c) *SFDA methods* are generally based on **white-box** source models, which are used as baselines to make comparisons. To be specific, we choose the following methods as our counterparts:

- URMA (S & Fleuret, 2021), which requires a set of auxiliary segmentation decoders to reduce model uncertainty.
- HCL (Huang et al., 2021), which applies historical contrastive learning to avoid overfitting pseudo-labeled data.
- LD (You et al., 2021), which introduces negative learning to mitigate the noise in pseudo-labels.
- SFUDA (Ye et al., 2021), which constructs a virtual data domain as a proxy of the source domain to regularize model adaptation.

Table 2 Quantitative results on *GTA5* → *Cityscapes* Benchmark

Method	Road	Sidewk.	Build	Wall	Fence	Pole	Light	Sign	Vege	Terr	Sky	Person	Rider	Car	Truck	Bus	Train	Motor	Bike	**mIoU**
Source Only (Tsai et al., 2018)	75.8	16.8	77.2	12.5	21.0	25.5	30.1	20.1	81.3	24.6	70.3	53.8	26.4	49.9	17.2	25.9	6.5	25.3	36.0	36.6
Source Only* (Kundu et al., 2021)	85.8	36.5	81.8	24.8	22.4	29.0	30.3	21.0	83.6	36.1	79.9	56.7	26.6	83.7	36.3	41.4	0.0	20.6	22.9	43.1
UDA methods																				
AdvEnt (Vu et al., 2019)	89.9	36.5	81.6	29.2	25.2	28.5	32.3	22.4	83.9	34.0	77.1	57.4	27.9	83.7	29.4	39.1	1.5	28.4	23.3	43.8
CRST (Zou et al., 2019)	91.0	55.4	80.0	33.7	21.4	37.3	32.9	24.5	85.0	34.1	80.8	57.7	24.6	84.1	27.8	30.1	26.9	26.0	42.3	47.1
DAST (Yu et al., 2021)	92.2	49.0	84.3	36.5	28.9	33.9	38.8	28.4	84.9	41.6	83.2	60.0	28.7	87.2	45.0	45.3	7.4	33.8	32.8	49.6
BCDM (Li et al., 2021)	90.5	37.3	83.7	39.2	22.2	28.5	36.0	17.0	84.2	35.9	85.8	59.1	35.5	85.2	31.1	39.3	21.1	26.7	27.5	46.6
MRNeT+ (Zheng & Yang, 2021)	90.4	31.2	85.1	36.9	25.6	37.5	48.8	48.5	85.3	34.8	81.1	64.4	36.8	86.3	34.9	52.2	1.7	29.0	44.6	50.3
DACS (Tranheden et al., 2021)	89.9	39.7	87.9	30.7	39.5	38.5	46.4	52.8	88.0	44.0	88.8	67.2	35.8	84.5	45.7	50.2	0.0	27.3	34.0	52.1
ProDA (Zhang et al., 2021)	87.8	56.0	79.7	46.3	44.8	45.6	53.5	53.5	88.6	45.2	82.1	70.7	39.2	88.8	45.5	59.4	1.0	48.9	56.4	57.5
CPSL (Li et al., 2022)	92.3	59.9	84.9	45.7	29.7	52.8	61.5	59.5	87.9	41.5	85.0	73.0	35.5	90.4	48.7	73.9	26.3	53.8	53.9	60.8
DAFormer (Hoyer et al., 2022)	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
White-Box SFDA methods																				
URMA (S & Fleuret, 2021)	92.3	55.2	81.6	30.8	18.8	37.1	17.7	12.1	84.2	35.9	83.8	57.7	24.1	81.7	27.5	44.3	6.9	24.1	40.4	45.1
HCL (Huang et al., 2021)	92.0	55.0	80.4	33.5	24.6	37.1	35.1	28.8	83.0	37.6	82.3	59.4	27.6	83.6	32.3	36.6	14.1	28.7	43.0	48.1
LD (You et al., 2021)	91.6	53.2	80.6	36.6	14.2	26.4	31.6	22.7	83.1	42.1	79.3	57.3	26.6	82.1	41.0	50.1	0.3	25.9	19.5	45.5
SFUDA (Ye et al., 2021)	95.2	40.6	85.2	30.6	26.1	35.8	34.7	32.8	85.3	41.7	79.5	61.0	28.2	86.5	41.2	45.3	15.6	33.1	40.0	49.4
GtA (Kundu et al., 2021)	91.4	52.9	85.3	37.3	31.3	36.4	37.7	35.1	86.2	48.2	89.6	61.9	34.2	86.8	51.1	50.8	3.9	42.4	53.5	53.5
ATP1 (Wang et al., 2022)	93.2	55.8	86.5	45.2	27.3	36.6	42.8	37.9	86.0	43.1	87.9	63.5	15.3	85.5	41.2	55.7	0.0	38.1	57.4	52.6
CROTS*	92.2	55.2	86.0	40.4	31.5	36.8	39.5	36.8	86.3	45.7	90.7	62.7	35.3	87.0	52.4	52.1	0.2	43.9	55.0	<u>54.2</u>
Black-Box SFDA methods																				
ATP2 (Wang et al., 2022)	83.6	25.8	81.9	30.2	25.2	27.9	36.2	28.7	84.8	34.4	77.5	62.2	35.7	81.5	32.3	16.8	0.0	41.7	53.5	45.3
CROTS	87.6	42.6	85.7	22.5	38.9	38.3	50.3	56.2	88.0	38.8	90.7	68.3	38.7	83.5	20.7	40.3	11.7	42.7	60.6	53.0
CROTS*	92.0	52.4	85.9	37.3	35.8	34.6	42.2	38.4	86.9	45.6	91.1	65.1	36.1	87.3	41.6	51.1	0.0	41.4	56.2	**53.7**

CROST adapts the source model in (Tsai et al., 2018) whereas CROST* is based on the source model in Kundu et al. (2021). The ***mIoU*** column shows the average performance of all the classes, where the best performance for white-box SFDA methods is shown with orange underline, and the best performance for black-box SFDA is shown in blue bold

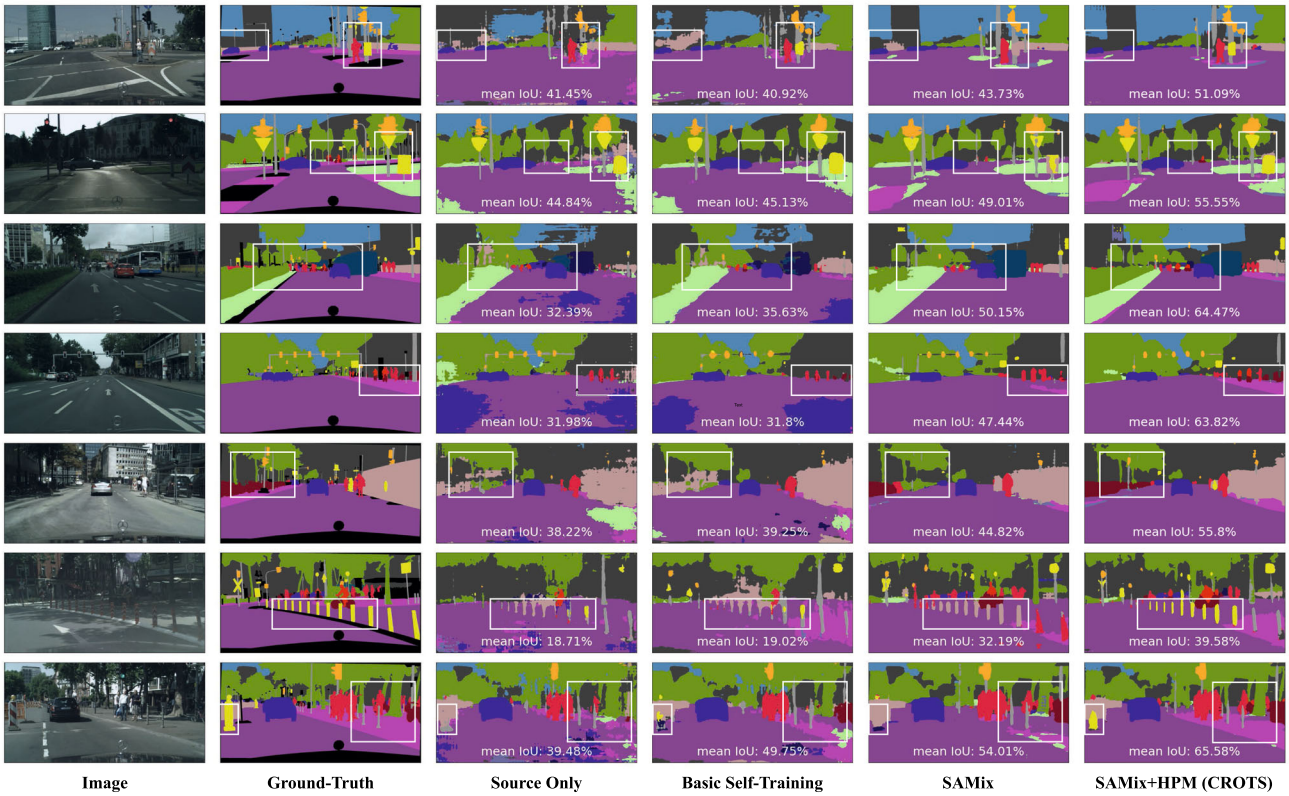

Fig. 7 Qualitative results on the *GTA5* → *Cityscapes* Benchmark

- GtA (Kundu et al., 2021), which trains a domain generalized source model and employs self-training to adapt it towards target data.
- ATP (Wang et al., 2022), which adapts pre-trained source models in three steps: implicit feature alignment, bidirectional self-training, and information propagation.

(d) *Black-Box SFDA methods* are different variants of our CROTS, which employ different black-box pre-trained source models for target adaptation. Black-Box setting is much less studied in the field of SFDA. Among the literature, ATP (Wang et al., 2022) provides its performance under the black-box setting, which is also chosen as our black-box counterpart.

4.2 Evaluation Results

To answer **RQ1**, we apply *CROTS* in the domain adaptive segmentation tasks towards the Cityscapes dataset. We compare the performance with leading SFDA counterparts and report the performance as follows:

4.2.1 Results on the GTA5 Benchmark

Table 2 reports the quantitative results of cross-domain semantic segmentation on the *GTA5* → *Cityscapes* benchmark. The results validate the superiority of the proposed method. In the table, *CROTS* denotes applying our method to adapt the pre-trained source model provided by AdaptSegNet (Tsai et al., 2018), whereas *CROTS** denotes our method with the pre-trained model from GtA (Kundu et al., 2021). Empirical results indicate that the target model trained with the CROTS framework achieves the best mean IoU of *54.2%* under the white-box SFDA setting. Under the black-box SFDA setting, the proposed CROTS still achieves a 53.7% mIoU, even outperforming many leading white-box SFDA methods that require the existence of pre-trained weights. Specifically, compared with GtA (Kundu et al., 2021), *CROTS* consistently improves the performance for most of the classes.

Figure 7 delivers the qualitative results of the proposed method. With the proposed CROTS, we effectively improve the noisy prediction of source-only models, providing better predictions. In Fig. 7, the region highlighted by the white rectangle denotes the detailed comparison. The results also validate the effectiveness of the *Spatial-Aware Data Mixing (SAMix)* augmentation and the *Rare-Class-Patches Mining (RCPM)* regularization, especially for long-tailed classes and the boundary of objects.

4.2.2 Results on the SYNTHIA Benchmark

Table 3 reports the results on the *SYNTHIA* → *Cityscapes* benchmark, which again proves the effectiveness of our proposed method, achieving the best mean IoU of 60.3% and 59.3% under white-box and black-box SFDA settings, respectively.

The above results confirm the superiority of *CROTS* and give a positive answer to **RQ1**: *CROTS* well tackles the SFDA problem, achieving superior performance to leading SFDA methods.

Besides, for state-of-the-art UDA methods (e.g., CPSL(Li et al., 2022)), as mentioned in Sect. 3, these methods may fail to keep their leading performance under the source-free setting. As a contrast, despite the lack of source data, the proposed *CROTS* works well, even outperforming typical UDA counterparts (e.g., DAST(Yu et al., 2021), MRNet(Zheng & Yang, 2021)).

4.3 Effectiveness of SAMix

In order to answer **RQ2**, we conduct ablation studies on the *GTA5* → *Cityscapes* benchmark to investigate the contribution of Spatial-Aware Data Mixing. The ablation results are reported in Table 4.

Since data mixing has been widely adopted as data augmentation in semi-supervised semantic segmentation, to validate the effectiveness of the proposed Spatial-Aware Data Mixing, we compare its performance of different mixing strategies:

- *CutMix* (Yun et al., 2019) mixes two samples by a rectangle mask, the performance of which is limited by the number of mixing samples.
- *CowMix* (French et al., 2020b) mixes the given two images by random image masks. The pseudo-labels can be extremely noisy considering the domain discrepancy, making the mixed samples meaningless. Besides, randomly mixing small patches also involves more noisy predictions. The severe noise makes CowMix unable to overcome the overfitting issue, suffering decreasing performance after two epochs. Therefore, CowMix is unsuitable for the SFDA task.
- *ClassMix* (Olsson et al., 2021) mixes two samples by randomly sampling class masks (calculated from pseudo-labels). Considering the cross-domain discrepancy, the masks might be noisy, leading to limited performance gain.

Figure 8 plots the validation mIoU during the training process of the first round of self-training. The descending lines mean that the target model begins overfitting noisy pseudo-labels. As shown in the figure, CowMix suffers from a descending curve, which proves that CowMix is not suitable for the SFDA task. Among *previous* mixing strategies, ClassMix achieves the best performance. As shown in the figure, Spatial-Aware Data Mixing further improve ClassMix to fit the task of SFDA.

The superiority of Spatial-Aware Data Mixing can be attributed to two aspects: (a) Mixing different images leads to diverse statistical data, which increases the data diversity and updates the Batch-Norm statistics so that they can approximate a wider range of distribution. (b) Keeping the spatial prior of classes and patches leads to reasonable sample generation, which also contributes to contextual consistency (as demonstrated in Sect. 1).

In Fig. 8, by comparing *SAMix* and *SAMix w/o spatial prior*, we validate the necessity of regularizing spatial prior during mixing samples. Meanwhile, the sharp contrast between *SAMix* and *SAMix w/o random masks* shows that it is crucial to introduce randomization for mixing.

Overall, Fig. 8 demonstrates that the proposed *Spatial-Aware Data Mixing* achieves the best performance improvement for SFDA, outperforming other mixing strategies. Besides, the task model might overfit the pseudo-labeled target data as the training process goes on, leading to descending performance curves. For example, although CowMix increases the data diversity, it cannot tackle the overfitting issue, which even leads to under-fitting results. In contrast, the proposed *Spatial-Aware Data Mixing* is capable of maintaining its performance even in the later stage of training. The quantitative results in Table 4 also prove the superiority of Spatial-Aware Data Mixing. After selecting the optimal hyper-parameters ($\beta = 0.25$ and $N = 4$), Spatial-Aware Data Mixing beats the other mixing strategies.

The above analysis gives a positive answer to **RQ2**: Spatial-Aware Data Mixing helps enrich the diversity of target pseudo-labeled data and mitigate the overfitting issue.

4.4 Effectiveness of RCPM

In order to answer **RQ3**, we conduct ablation studies on the *GTA5* → *Cityscapes* benchmark to investigate the contribution of Rare-Class Patches Mining. The ablation results are reported in Table 5.

Rare-Class Patches Mining aims to tackle the challenge of class imbalance. As listed in Table 4, without extra regularization, despite the promising performance improvement, only applying mixing strategies may lead to sub-optimal performance for long-tailed rare classes. For instance, in Table 4, all methods can hardly segment the *train* class, leading to an mIoU of 0.0%.

Table 5 demonstrates that the proposed *Rare-Class Patches Mining* mechanism further improves the segmentation results, especially over those hard-to-adapt long-tailed classes, e.g., *train, motorcycle, rider, light*, etc. The improvement proves

Table 3 Quantitative results on *SYNTHIA* → *Cityscapes* Benchmark

Method	Road	Sidewk.	Build	Wall*	Fence*	Pole*	Light	Sign	Vege	Sky	Person	Rider	Car	Bus	Motor	Bike	**mIoU**	**mIoU***
Source Only (Tsai et al., 2018)	47.1	23.3	75.5	7.1	0.1	23.9	5.1	9.2	74.0	73.5	51.1	20.9	39.1	17.7	18.4	34.0	32.5	37.6
Source Only* (Kundu et al., 2021)	87.2	35.0	77.3	5.8	0.4	29.5	15.6	18.3	75.0	78.3	52.6	21.5	76.2	29.3	13.7	26.8	40.1	46.7
UDA methods																		
AdvEnt (Vu et al., 2019)	87.0	44.1	79.7	9.6	0.6	24.3	4.8	7.2	80.1	83.6	56.4	23.7	72.7	32.6	12.8	33.7	40.8	47.6
CRST (Zou et al., 2019)	67.7	32.2	73.9	10.7	1.6	37.4	22.2	31.2	80.8	80.5	60.8	29.1	82.8	25.0	19.4	45.3	43.8	50.1
DAST (Yu et al., 2021)	87.1	44.5	82.3	10.7	0.8	29.9	13.9	13.1	81.6	86.0	60.3	25.1	83.1	40.1	24.4	40.5	45.2	52.5
MRNet+ (Zheng & Yang, 2021)	87.6	41.9	83.1	14.7	1.7	36.2	31.3	19.9	81.6	80.6	63.0	21.8	86.2	40.7	23.6	53.1	47.9	54.9
DACS (Tranheden et al., 2021)	80.6	25.1	81.9	21.5	2.9	37.2	22.7	24.0	83.7	90.8	67.6	38.3	82.9	38.9	28.5	47.6	48.3	54.8
ProDA (Zhang et al., 2021)	87.8	45.7	84.6	37.1	0.6	44.0	54.6	37.0	88.1	84.4	74.2	24.3	88.2	51.1	40.5	45.6	55.5	62.0
CPSL (Li et al., 2022)	87.2	43.9	85.5	33.6	0.3	47.7	57.4	37.2	87.8	88.5	79.0	32.0	90.6	49.4	50.8	59.8	57.9	65.3
DAFormer (Hoyer et al., 2022)	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	67.4
White-Box SFDA methods																		
URMA (S & Fleuret, 2021)	59.3	24.6	77.0	14.0	1.8	31.5	18.3	32.0	83.1	80.4	46.3	17.8	76.7	17.0	18.5	34.6	39.6	45.0
HCL (Huang et al., 2021)	80.9	34.9	76.7	6.6	0.2	36.1	20.1	28.2	79.1	83.1	55.6	25.6	78.8	32.7	24.1	32.7	43.5	50.2
LD (You et al., 2021)	77.1	33.4	79.4	5.8	0.5	23.7	5.2	13.0	81.8	78.3	56.1	21.6	80.3	49.6	28.0	48.1	42.6	50.1
SFUDA (Ye et al., 2021)	90.9	45.5	80.8	3.6	0.5	28.6	8.5	26.1	83.4	83.6	55.2	25.0	79.5	32.8	20.2	43.9	44.2	51.9
GtA (Kundu et al., 2021)	90.5	50.0	81.6	13.3	2.8	34.7	25.7	33.1	83.8	89.2	66.0	34.9	85.3	53.4	46.1	46.6	52.0	60.1
ATP (Wang et al., 2022)	90.1	46.3	82.5	0.0	0.1	31.7	10.7	17.9	85.1	87.7	64.6	34.6	86.4	54.8	33.7	58.3	49.0	57.9
CROTS*	89.5	51.2	83.1	14.6	4.3	35.1	32.5	35.4	85.5	80.4	67.8	35.4	81.7	45.2	47.5	48.9	52.4	60.3
Black-Box SFDA methods																		
CROTS	80.4	44.2	83.1	10.7	0.1	28.8	31.0	25.5	85.1	88.7	61.3	22.6	58.6	42.4	24.2	50.9	45.8	53.7
CROTS*	89.4	41.6	82.7	15.1	1.2	34.7	33.7	25.7	83.7	87.9	66.6	34.6	85.4	45.9	43.5	49.6	51.3	**59.3**

CROST adapts the source model in (Tsai et al., 2018) *whereas CROST* are based on the source model in* Kundu et al. (2021). The ***mIoU*** column shows the average performance of all the classes, and the ***mIoU**** *column is the average performance that excludes three classes marked with *, where the best performance for white-box SFDA methods is shown with* orange underline, and the best performance for black-box SFDA is shown in blue bold

Table 4 Ablation of spatial-aware data mixing on textitGTA5 → *Cityscapes* Benchmark

Method	Road	Sidewk.	Build	Wall	Fence	Pole	Light	Sign	Vege	Terr	Sky	Person	Rider	Car	Truck	Bus	Train	Motor	Bike	**mIoU**
Previous augmentation strategies																				
ClassMix (Olsson et al., 2021)	87.4	13.2	87.9	25.1	31.2	36.8	46.9	56.1	88.3	47.3	89.8	64.5	19.9	88.4	42.0	47.0	0.0	0.0	57.6	48.9
CowMix (French et al., 2020b)	87.8	19.9	83.1	25.1	32.2	32.7	35.5	50.3	84.5	41.2	85.2	60.5	5.5	84.6	31.1	42.2	0.0	0.0	55.2	45.1
CutMix (Yun et al., 2019)	89.9	34.9	84.8	24.3	36.7	34.4	44.0	46.6	84.6	30.2	88.3	57.9	25.8	85.4	30.6	44.0	0.0	23.7	48.9	48.2
Spatial-aware data mixing (SAMix) with different mixing prior ($N = 4$)																				
SAMix (w/o random masks)	87.9	41.1	84.4	14.2	38.0	26.1	37.2	36.5	84.2	44.0	87.3	58.9	33.3	73.2	26.9	37.7	0.0	30.9	51.8	47.0
SAMix ($\beta = 0.1$)	90.7	43.6	85.2	28.9	37.7	30.1	42.0	45.1	83.8	33.6	89.5	56.6	14.6	85.3	35.1	42.0	0.0	29.0	53.3	48.8
SAMix ($\beta = 0.25$)	91.8	47.1	85.0	27.5	36.7	31.9	42.5	47.6	83.9	34.1	88.6	56.8	23.1	85.1	48.4	48.2	0.0	26.6	53.5	**50.4**
SAMix ($\beta = 0.5$)	91.5	52.4	84.8	26.8	39.4	30.7	42.7	52.8	86.1	39.9	88.6	57.4	33.0	82.5	31.1	51.7	0.0	0.0	53.3	49.7
SAMix ($\beta = 0.75$)	90.5	41.1	84.4	7.6	38.9	31.3	40.6	45.9	85.6	46.0	86.5	58.7	30.1	83.4	25.0	39.3	0.0	23.5	55.7	48.1
SAMix ($\beta = 1.0$)	90.2	43.6	84.9	17.0	35.5	32.0	40.0	42.5	85.5	42.2	88.4	57.8	26.5	85.3	37.0	43.1	0.0	0.0	52.3	47.6
Spatial-aware data mixing (SAMix) with different numbers of patches ($\beta = 0.25$)																				
SAMix ($N = 1$)	89.7	41.8	85.4	26.6	29.2	35.0	39.3	51.5	86.9	42.0	89.2	60.3	25.4	83.6	26.3	36.5	0.0	31.7	52.4	49.1
SAMix ($N = 2$)	90.5	40.2	85.8	14.0	35.7	33.5	39.6	51.3	86.7	41.5	88.5	59.0	28.7	84.5	32.2	47.7	0.0	26.6	49.5	49.2
SAMix ($N = 3$)	92.0	51.0	84.3	0.0	37.7	32.5	40.8	47.9	86.6	40.9	88.3	61.9	31.3	85.8	44.4	44.8	0.0	25.1	50.8	49.8
SAMix ($N = 4$)	91.8	47.1	85.0	27.5	36.7	31.9	42.5	47.6	83.9	34.1	88.6	56.8	23.1	85.1	48.4	48.2	0.0	26.6	53.5	**50.4**
SAMix ($N = 6$)	91.2	49.9	83.7	0.0	38.2	30.5	39.1	48.8	86.6	38.5	88.6	61.0	31.2	85.7	35.6	41.5	0.0	25.0	49.1	48.7
SAMix ($N = 9$)	91.2	48.2	84.3	0.0	35.8	31.7	40.4	47.6	87.1	42.8	88.6	60.1	17.5	84.9	38.7	44.2	0.0	0.0	45.5	46.8

The **mIoU** column shows the average performance of all the classes, where the best performance is shown in blue bold

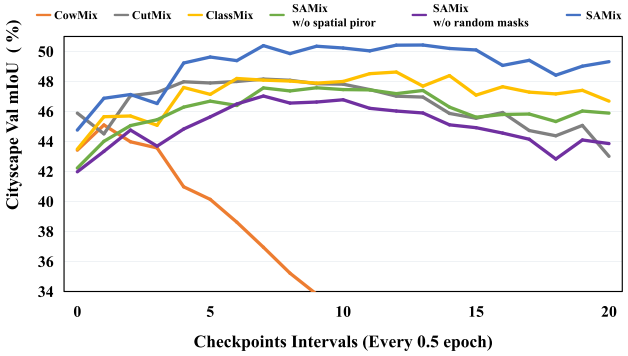

Fig. 8 Comparison of different mixing strategies. For *CowMix*, since its performance drops significantly after several epochs, we interrupt its training and report the mIoU before early-stopping. *SAMix w/o spatial prior* extends *CutMix* by increasing the number of mixed samples. Compared with *SAMix w/o random mask*, *SAMix* injects randomization into the mixing masks

the effectiveness of *Rare-Class Patches Mining* to mitigate the class imbalance. In Table 4, applying *Rare-Class Patches Mining* consistently improves different mixing strategies. Moreover, when combined with the proposed *Rare-Class Patches Mining*, the improvements are more significant.

These results give a positive answer to **RQ3**: Rare-Class Patches Mining helps increase the contribution of long-tailed classes and alleviates the issue of class imbalance.

However, it should be noted that the aim of RCPM is to maintain the performance of long-tailed rare classes. For classes that the original pre-trained source model fails to segment, Rare-Class Patches Mining may also fail to take effect. For example, as shown in Table 2, for black-box SFDA, CROTS* (with stronger pre-trained source model) generally performs better (+ 0.3% mIoU) than CROTS. But its performance on the "train" class is much less than CROTS (0.0% vs. 11.7%). This can be explained by class-wise performance gap among different pre-trained source models.

Besides, the performance drop for the main classes (e.g., "road") is a trade-off between the rare classes and the main classes. The experimental results in Table 5 show that RCPM can improve the overall performance (from 50.4% mIoU to 53.0% mIoU), which improves significantly for rare classes(e.g., for the "train" class, improving the mIoU from 0.0 to 11.7%). With this consideration, the slight performance degradation on the main classes (e.g., −4.2% mIoU of the "road" class) is allowed in RCPM.

4.5 Parameter Tuning

In order to answer **RQ4**, we experiment with different hyper-parameters in CROTS.

Table 5 Ablation of rare-class patches mining on *GTA5* → *Cityscapes* Benchmark

Method	Road	Sidewk.	Build	Wall	Fence	Pole	Light	Sign	Vege	Terr	Sky	Person	Rider	Car	Truck	Bus	Train	Motor	Bike	**mIoU**
ClassMix (Olsson et al., 2021)	87.4	13.2	87.9	25.1	31.2	36.8	46.9	56.1	88.3	47.3	89.8	64.5	19.9	88.4	42.0	47.0	0.0	0.0	57.6	48.9
+RCPM	82.1	32.6	86.4	24.2	34.1	36.4	44.2	52.8	87.9	26.0	88.8	64.1	35.7	85.2	32.4	44.1	0.0	41.3	59.6	**50.4**
CutMix (Yun et al., 2019)	89.9	34.9	84.8	24.3	36.7	34.4	44.0	46.6	84.6	30.2	88.3	57.9	25.8	85.4	30.6	44.0	0.0	23.7	48.9	48.2
+RCPM	88.9	33.6	83.8	23.1	30.7	34.4	40.0	45.0	87.0	39.2	87.4	61.7	32.5	84.4	27.7	44.0	0.0	40.2	56.6	**49.5**
SAMix	91.8	47.1	85.0	27.5	36.7	31.9	42.5	47.6	83.9	34.1	88.6	56.8	23.1	85.1	48.4	48.2	0.0	26.6	53.5	50.4
+RCPM (size = 32)	86.0	33.3	83.3	12.6	34.6	33.8	43.6	46.1	87.2	38.5	88.8	61.5	36.8	82.7	23.0	42.7	3.2	41.7	56.1	49.2
+RCPM (size = 64)	87.6	42.6	85.7	22.5	38.9	38.3	50.3	56.2	88.0	38.8	90.7	68.3	38.7	83.5	20.7	40.3	11.7	42.7	60.6	**53.0**
+RCPM (size = 128)	88.9	41.5	83.8	23.1	32.6	34.4	40.0	45.0	87.0	39.2	87.4	61.7	32.5	84.4	47.3	44.0	8.1	40.2	56.6	51.5

4.5.1 β in SAMix

As mentioned in Sect. 3.2, we sample the mixing masks from a *Beta* distribution. Following previous studies of mixing strategies(Zhang et al., 2018), we set the α and β parameters to be the same, uniformed as β. To obtain optimal performance, we investigate different hyperparameters that control the *Beta* distribution during experiments. Table 4 reports the tuning results, where *SAMix (w/o random masks)* denotes that a target image is uniformly mixed from the fixed part of other images without randomization. The results show that Spatial-Aware Data Mixing consistently improves performance. The optimal value of β is 0.25. The comparison between *SAMix (w/o random masks)* and the other variants shows that Spatial-Aware Data Mixing introduces more diversity for target data by injecting randomization into mixing masks, which leads to superior performance to that of fixed-region mixing.

4.5.2 N in SAMix

The proposed Spatial-Aware Data Mixing extends the CutMix (Yun et al., 2019) method by increasing the number of mixing samples and regularizing the spatial prior. To investigate the necessity of increasing the number of mixing samples, we conduct experimental studies to compare different Spatial-Aware Data Mixing variants.

As demonstrated in Table 4, with increasing mixing patches, the performance is progressively improved. It is observed that when N goes beyond 4, the performance gain is limited. When splitting the whole image into too many patches (e.g., $N = 9$), the performance even gets much worse. This is caused by the violation of spatial prior. As stated in the manuscript, one important improvement of SAMix is the employment of spatial prior, when splitting the image into too many patches, the mixed sample will have less consistent spatial layouts, leading to worse performance.

4.5.3 Size of Rare-Class Patches

we also conduct extra experiments to validate the performance of RCPM under different patch sizes. The results are listed in Table. 5.

- With smaller patches (32×32), RCPM ignores the integrity of rare-class objects. For example, only a small part of a "bicycle" patch is used as a rare-class patch. This makes the generated patches less likely to be correctly segmented, leading to worse performance than that with 64×64 patches (with mIoU 49.2 vs. 53.0%).
- With larger patches (128×128), RCPM works similarly to CutMix. However, with the increased size of patches, since the patch is dominated by the main classes, the contribution of rare-class pixels is weakened. In this way, RCPM is less effective to alleviate the class-imbalance issue. During experiments, the overall performance of larger patches is similar to that with 64×64 patches (with mIoU 51.5 vs. 53.0%). However, the improvements in long-tailed rare classes are less significant. For example, the IoU of "train" class decreases from 11.7 to 8.1%.

The above results answer **RQ4**: when performing SFDA tasks, Spatial-Aware Data Mixing relies on the hyperparameter β and N. Considering the practical experimental environments, the best value of β is set to 0.25, which injects randomness into mixing masks and achieves optimal performance. The best value for N is set to 4, which effectively increases the diversity of mixing samples and mitigates the overfitting issue. Besides, the best size of rare-class patches is set to be 64×64.

4.6 Performance under the Black-Box Setting

With respect to **RQ5**, we extend CROTS to adapt the black-box pre-trained model. The results are given in Tables 2 and 3. CROTS outperforms previous black-box SFDA methods (e.g., ATP (Wang et al., 2022)). Even under the black-box setting, CROTS achieves superior performance to many leading white-box counterparts, which confirms its capability to adapt black-box source models and answers **RQ5**.

5 Conclusion

This study proposes a novel teacher–student learning framework for source-free domain adaptive (SFDA) semantic segmentation.

(1) To tackle the issue of overfitting noisy pseudo-labeled target data, we introduce the intra-domain teacher model to enrich the diversity of target data, leading to an augmentation strategy named Spatial-Aware Data Mixing, which combines different target samples to synthesize new target data, enriching the diversity of training data and alleviating the overfitting issue.
(2) Aiming at the class-imbalance issue, we analyzed the class-wise performance difference between pre-trained source models and adapted target models. Since pre-trained source models can perform better for long-tailed rare classes, we introduce pre-trained source models as the inter-domain teacher model, which contributes to a training regularization mechanism termed Rare-Class Patches Mining. Rare-Class Patches Mining increases the contribution of the minority classes, which regularizes the task model to avoid forgetting those rare classes, mitigating the class imbalance issue.

(3) The cooperation between Spatial-Aware Data Mixing and Rare-Class Patches Mining results in the proposed CROTS, a cross-domain teacher–student learning framework for source-free domain adaptive semantic segmentation.

Experimental results verify that our proposed CROTS effectively tackles the SFDA problem, outperforming many leading SFDA counterparts. Additionally, we extend CROTS for black-box SFDA. Extensive experiments show that the proposed CROTS is capable of adapting black-box source models and even achieves superior performance to many white-box SFDA methods.

Acknowledgements We extend our gratitude to Mengzhu Wang and Xifeng Guo, who have given us valuable suggestions to improve our manuscript.

Author Contributions XL and WC made substantial contributions to the conception or design of the work. ZL, LY, SW, SW, and CL made contributions to the acquisition, analysis, or interpretation of data. All the authors drafted the work or revised it critically. All the authors approved the version to be published.

Funding This work was supported by the Natural Science Foundation of Hunan Province of China (No. 2022JJ30666), the Independent and Open Subject Fund from State Key Laboratory of High Performance Computing, National University of Defense Technology (No. 202101-10), and the National Key Technologies Research and Development Program of China (No. 2018YFB0204301).

Availability of data and materials The datasets used in this manuscript can be downloaded publicly from the official websites.

Declarations

Conflicts of interest The authors have no financial or proprietary interests in any material discussed in this article.

Ethical approval Not applicable.

Consent to participate Not applicable.

Consent for publication The authors confirm that: (1) the work described has not been published before; (2) the manuscript is not under consideration for publication elsewhere; (3) the publication has been approved by all co-authors; (4) the publication has been approved by the responsible authorities at the institution where the work is carried out.

Code availability The codes used in this manuscript will be made publicly available.

References

Ahmed, S.M., Raychaudhuri, D.S., Paul, S., Oymak, S., & Roy-Chowdhury A.K. (2021). Unsupervised multi-source domain adaptation without access to source data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10098–10107).

Asgari Taghanaki, S., Abhishek, K., Cohen, J. P., Cohen-Adad, J., & Hamarneh, G. (2021). Deep semantic segmentation of natural and medical images: A review. *Artificial Intelligence Review, 54*(1), 137–178.

Bochkovskiy, A., Wang, C.-Y., & Liao, H.-Y.M. (2020). *Yolov4: Optimal speed and accuracy of object detection*. arXiv preprint arXiv:2004.10934 .

Chang, W.-L., Wang, H.-P., Peng, W.-H., & Chiu, W.-C. (2019). All about structure: Adapting structural information across domains for boosting semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 1900–1909).

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2017). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern analysis and Machine Intelligence, 40*(4), 834–848.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., & Schiele, B. (2016). The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 3213–3223).

Du, Z., Li, J., Su, H., Zhu, L., & Lu, K. (2021). Crossdomain gradient discrepancy minimization for unsupervised domain adaptation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 3937–3946).

French, G., Laine, S., Aila, T., Mackiewicz, M., & Finlayson, G.D. (2020a). Semi-supervised semantic segmentation needs strong, varied perturbations. In *Proceedings of the British machine vision conference*. BMVA Press.

French, G., Laine, S., Aila, T., Mackiewicz, M., & Finlayson, G.D. (2020b). Semi-supervised semantic segmentation needs strong, varied perturbations. In *Proceedings of the 31st British machine vision conference*.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., & Lempitsky, V. (2016). Domain-adversarial training of neural networks. *The Journal of Machine Learning Research, 17*(1), 2096.

Gou, J., Yu, B., Maybank, S. J., & Tao, D. (2021). Knowledge distillation: A survey. *International Journal of Computer Vision, 129*(6), 1789–1819.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 770–778). Las Vegas, NV, USA: IEEE. https://doi.org/10.1109/CVPR.2016.90

Hoyer, L., Dai, D., & Gool, L.V. (2022). DAFormer: Improving Network Architectures and Training Strategies for Domain-Adaptive Semantic Segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 9924–9935).

Hu, X., Tang, C., Chen, H., Li, X., Li, J., & Zhang, Z. (2022). Improving image segmentation with boundary patch refinement. *International Journal of Computer Vision, 130*(11), 2571–2589.

Huang, J., Guan, D., Xiao, A., & Lu, S. (2021). Model adaptation: Historical contrastive learning for unsupervised domain adaptation without source data. In *Advances in neural information processing systems* (Vol. 34, pp. 3635–3649).

Kamann, C., & Rother, C. (2021). Benchmarking the robustness of semantic segmentation models with respect to common corruptions. *International Journal of Computer Vision, 129*(2), 462–483.

Kundu, J.N., Kulkarni, A., Singh, A., Jampani, V., & Babu, R.V. (2021). Generalize then Adapt: Source-Free Domain Adaptive Semantic Segmentation. In *Proceedings of the IEEE/CVF International conference on computer vision* (pp. 7026–7036).

Kurmi, V.K., Subramanian, V.K., & Namboodiri, V.P. (2021). Domain impression: A source data free domain adaptation method. In *Proceedings of the IEEE winter conference on applications of computer vision* (pp. 615–625).

Lee, C.-Y., Batra, T., Baig, M.H., & Ulbricht, D. (2019). Sliced wasserstein discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10285–10295).

Li, H., Wan, R., Wang, S., & Kot, A. C. (2021). Unsupervised domain adaptation in the wild via disentangling representation learning. *International Journal of Computer Vision, 129*(2), 267–283. https://doi.org/10.1007/s11263-020-01364-5

Li, R., Jiao, Q., Cao, W., Wong, H.-S., & Wu, S. (2020). Model adaptation: Unsupervised domain adaptation without source data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 9638–9647).

Li, R., Li, S., He, C., Zhang, Y., Jia, X., & Zhang, L. (2022). Class-balanced pixel-level self-labeling for domain adaptive semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.

Li, S., Lv, F., Xie, B., Liu, C.H., Liang, J., & Qin, C. (2021). Bi-classifier determinacy maximization for unsupervised domain adaptation. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 35, pp. 8455–8464).

Li, Y., Yuan, L., & Vasconcelos, N. (2019). Bidirectional learning for domain adaptation of semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 6936–6945).

Liang, J., Hu, D., & Feng, J. (2020). Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *Proceedings of the 37th international conference on machine learning* (pp. 6028–6039). PMLR.

Liang, J., Hu, D., Jiashi, F., & He, R. (2022). Dine: Domain adaptation from single and multiple black-box predictors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.

Liu, Y., Tian, Y., Chen, Y., Liu, F., Belagiannis, V., & Carneiro, G. (2022). Perturbed and strict mean teachers for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (p. 10).

Liu, Y., Zhang, W., & Wang, J. (2021). Source-free domain adaptation for semantic segmentation. In *Proceeedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 1215–1224).

Loshchilov, I., & Hutter, F. (2017). Sgdr: Stochastic gradient descent with warm restarts. In *Proceedings of the international conference on learning representations*.

Lu, Z., Yang, Y., Zhu, X., Liu, C., Song, Y.-Z., & Xiang, T. (2020). Stochastic classifiers for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 9111–9120).

Luo, Y., Zheng, L., Guan, T., Yu, J., & Yang, Y. (2019). Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 2507–2516).

Nath Kundu, J., Venkat, N., Rahul, M. V., & Venkatesh Babu, R. (2020). Universal sourcefree domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 4543–4552).

Olsson, V., Tranheden, W., Pinto, J., & Svensson, L. (2021). Classmix: Segmentation-based data augmentation for semi-supervised learning. In *Proceedings of the IEEE winter conference on applications of computer vision* (pp. 1368–1377).

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., & Chintala, S. (2019). Pytorch: An imperative style, highperformance deep learning library. In *Advances in neural information processing systems* (Vol. 32).

Richter, S. R., Vineet, V., Roth, S.,& Koltun, V. (2016). Playing for data: Ground truth from computer games. In *European conference on computer vision* (pp. 102–118).

Ros, G., Sellart, L., Materzynska, J., Vazquez, D., & Lopez, A.M. (2016). The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 3234–3243).

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., & Fei-Fei, L. (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision, 115*(3), 211–252.

Sivaprasad, P. T., & Fleuret, F. (2021). Uncertainty reduction for model adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 9608–9618).

Saito, K., Watanabe, K., Ushiku, Y., & Harada, T. (2018). Maximum classifier discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 3723–3732).

Summa, M.G., Bottou, L., Goldfarb, B., Murtagh, F., Pardoux, C., & Touati, M. (2010). Largescale machine learning with stochastic gradient descent léon bottou. In *Proceedings of the international conference on computational statistics* (pp. 33–42). Chapman and Hall/CRC. https://doi.org/10.1201/b11429-6

Tarvainen, A., & Valpola, H. (2017). Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in neural information processing systems* (Vol. 30). Curran Associates, Inc.

Tranheden, W., Olsson, V., Pinto, J., & Svensson, L. (2021). DACS: Domain adaptation via crossdomain mixed sampling. In *Proceedings of the IEEE winter conference on applications of computer vision* (pp. 1378–1388). IEEE. https://doi.org/10.1109/WACV48630.2021.00142

Tsai, Y.-H., Hung, W.-C., Schulter, S., Sohn, K., Yang, M.-H., & Chandraker, M. (2018). Learning to adapt structured output space for semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 7472–7481).

Vu, T. -H., Jain, H., Bucher, M., Cord, M., & Pérez, P. (2019). Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 2517–2526).

Wang, D., Shelhamer, E., Liu, S., Olshausen, B., & Darrell, T. (2021). Tent: Fully test-time adaptation by entropy minimization. In *Proceedings of the international conference on learning representations*.

Wang, Y., Liang, J., & Zhang, Z. (2022). Source data-free cross-domain semantic segmentation: Align, teach and propagate (No. arXiv:2106.11653)

Yang, S., Wang, Y., van de Weijer, J., Herranz, L., & Jui, S. (2021). Generalized sourcefree domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 8958–8967).

Yang, Y., Lao, D., Sundaramoorthi, G., & Soatto, S. (2020). Phase consistent ecological domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 9011–9020).

Ye, M., Zhang, J., Ouyang, J., & Yuan, D. (2021). Source data-free unsupervised domain adaptation for semantic segmentation. In *Proceedings of the 29th ACM international conference on multimedia* (p. 2233–2242).

You, F., Li, J., Zhu, L., Chen, Z., & Huang, Z. (2021). Domain adaptive semantic segmentation without source data. In *Proceedings of the 29th ACM international conference on multimedia* (pp. 3293–3302).

Yu, F., Zhang, M., Dong, H., Hu, S., Dong, B., & Zhang, L. (2021). Dast: Unsupervised domain adaptation in semantic segmentation based on discriminator attention and self-training. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 35, pp. 10754–10762).

Yu, L., Li, Z., Xu, M., Gao, Y., Luo, J., & Zhang, J. (2022). Distribution-aware margin calibration for semantic segmentation in images. *International Journal of Computer Vision, 130*(1), 95–110.

Yun, S., Han, D., Chun, S., Oh, S. J., Yoo, Y., Choe, J. (2019). Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision* (p. 6022–6031).

Zhang, H., Cisse, M., Dauphin, Y.N., & Lopez-Paz, D. (2018). Mixup: BEYOND EMPIRICAL RISK MINIMIZATION. In *Proceedings of the international conference on learning representations* (p. 13).

Zhang, H., Zhang, Y., Jia, K., & Lei, Z. (2021). Unsupervised domain adaptation of blackbox source models. In *Proceedings of the British machine vision conference*.

Zhang, P., Zhang, B., Zhang, T., Chen, D., Wang, Y., & Wen, F. (2021). Prototypical pseudo label denoising and target structure learning for domain adaptive semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 12409–12419).

Zhao, S., Li, B., Xu, P., Yue, X., Ding, G., & Keutzer, K. (2021). MADAN: multi-source adversarial domain aggregation network for domain adaptation. *International Journal of Computer Vision, 129*(8), 2399–2424.

Zheng, Z., & Yang, Y. (2021). Rectifying pseudo label learning via uncertainty estimation for domain adaptive semantic segmentation. *International Journal of Computer Vision, 129*(4), 1106–1120. https://doi.org/10.1007/s11263-020-01395-y

Zou, Y., Yu, Z., Kumar, B., & Wang, J. (2018). Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proceedings of the European conference on computer vision* (pp. 289–305).

Zou, Y., Yu, Z., Liu, X., Kumar, B., & Wang, J. (2019). Confidence regularized self-training. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 5982–5991).