Dynamic Mutual Adversarial Learning for Semi-supervised Semantic Segmentation of Underwater Images with Limited and Noisy Annotations

Han Chen , Ming Li , Yancheng Liu , Jingchun Zhou , Xianping Fu , Siyuan Liu * , Fei Richard Yu

Posted Date: 5 November 2025

doi: 10.20944/preprints202511.0348.v1

Keywords: underwater images; semantic segmentation; semi-supervised learning; dynamic mutual adversarial learning

 doi:10.20944/preprints202511.0348.v1

Article

Dynamic Mutual Adversarial Learning for Semi-supervised Semantic Segmentation of Underwater Images with Limited and Noisy Annotations

Han Chen [1], Ming Li [2], Yancheng Liu [1], Jingchun Zhou [3], Xianping Fu [3], Siyuan Liu [1,*] and Fei Richard Yu [2]

[1] Marine Engineering College, Dalian Maritime University, Dalian 116026, China
[2] Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ), Shenzhen 518000, China
[3] the College of Information Science and Technology, Dalian Maritime University, Dalian 116026, China
* Correspondence: dmu.s.y.liu@gmail.com; Tel.: +86-134-7899-6626

Abstract

Swift and accurate semantic segmentation of underwater images is key for precise object recognition in complex underwater environment. Nonetheless, the inherent complexity of these environments and the limited availability of labeled data pose significant challenges to underwater image segmentation. Traditional deep learning methods struggle to cope with limited and noisy annotations. In this paper, we delineate the formulation of a novel semi-supervised paradigm with dynamic mutual adversarial training for the semantic segmentation of underwater images. This paradigm identifies the sources of inaccuracies in pseudo-labeling by analyzing different confidence maps, generated by models with unique prior knowledge. A dynamic reweighing loss function is then employed to orchestrate the mutual instruction of two divergent models. Furthermore, the delineation of confidence map is facilitated via adversarial networks, which involves simultaneous adversarial refinement of the discrimination network and the segmentation model, using the predictions with high-confidence maps as pseudo-labels. Experimental results on public underwater datasets verify that the proposed method can effectively improve semantic segmentation performance under the condition of a small amount of labeled data.

Keywords: underwater images; semantic segmentation; semi-supervised learning; dynamic mutual adversarial learning

1. Introduction

Underwater environment perception, particularly optical visual perception, is crucial for the autonomous navigation and operation of underwater vehicles[1]. The timeliness and accuracy of semantic segmentation, a key technology in underwater perception, directly impact the overall performance of these systems. Therefore, efficient and precise semantic segmentation algorithms are essential for the practical application of intelligent sensing systems in underwater vehicles[2].

Deep neural networks, especially convolutional neural networks (CNNs), have recently shown great potential in the semantic segmentation of underwater images due to their excellent feature extraction and numerical regression capabilities[3,4]. However, most deep learning-based methods for underwater image segmentation require substantial amounts of labeled training data and precise pixel-wise labels for each image. The complex and unclear nature of underwater environments makes manual annotation labor-intensive and time-consuming, complicating the acquisition of large labeled datasets. To address this issue, semi-supervised semantic segmentation methods[5–8] have been proposed to reduce dependence on high-quality data by using self-/pseudo-supervision to perform

 doi:10.20944/preprints202511.0348.v1

segmentation without extensive manual labeling. Nonetheless, these methods often suffer from errors in pseudo labels. Previous self-training methods[9,10] always select pseudo labels based on confidence, derived from a model trained on a labeled subset. However, confidence-based methods have limitations: 1) To maintain a low error rate in pseudo labels, many low-confidence pseudo labels, which are naturally correct, are discarded; 2) Errors in pseudo-supervision from the model itself (i.e., self-error) can significantly hinder semi-supervised learning performance.

To overcome above challenges, we propose a novel semi-supervised method for semantic segmentation of underwater images, named Dynamic Mutual Adversarial Segmentation (DMAS). Our method employs a dual-model configuration where each model's segmentations serve as checks for the other to enable the identification and correction of mislabeling. The key features of the DMAS methodology are as follows:

1. The DMAS framework is based on two essential sub-processes: The first stage involves adversarial pre-training of two segmentation networks and their respective discriminators to develop two preliminary pseudo-label annotation models. The second stage is dynamic mutual learning, which measures the discrepancies between different segmentation models through confidence maps to mitigate the effects of potential pseudo label labeling errors, thereby enhancing the accuracy of the training process.
2. The adversarial training method is mainly used for training a segmentation model and a fully convolutional discriminator with labeled data to generate pseudo-labels. This allows the discriminator to differentiate between real label maps and their predicted counterparts by generating a confidence map. Also, it enables a quantitative assessment of segmentation accuracy in specific regions of the pseudo-labels.
3. Dynamic mutual learning guides different models based on their varying prior knowledge. It leverages the divergence between these models to detect inaccuracies in pseudo-label generation. By employing a dynamically reweighted loss function, it reflects the discrepancies between two models trained with each other's pseudo-labels, thereby assigning lower weights to pixels with a higher likelihood of error.
4. We validate the effectiveness of our proposed method on various underwater datasets, namely the DUT dataset and the SUIM dataset, demonstrating that the proposed semi-supervised learning algorithm is capable of enhancing the performance of models trained with limited and noisy annotations to be comparable to models fully-supervisedly trained with large amounts of labeled data.

2. Related Work

2.1. Pseudo-Label Methods for Semi-Supervised Semantic Segmentation

Pseudo-labeling[11,12] is a well-established technique in the domain of semi-supervised learning, with roots tracing back to some of the earliest works in the field[3]. Typically, pseudo-labeling based methods use models that have been trained on labeled datasets to predict provisional labels for unlabeled data. These predicted labels are then integrated into the existing dataset, which allows for the training of an improved model on this expanded dataset. Essentially, pseudo-labeling based methods can be divided into two categories: single-model based learning and mutual-model based learning.

Single-model based learning methods[13–16] rely on pseudo-labels generated by a single model within a self-training framework. However, these pseudo-labels are often unreliable since they rely solely on the prediction confidence of one model to filter out low-confidence labels. This strategy cannot detect and correct its labeling errors, which may result in the accumulation of bias and ultimately affect the training and segmentation effects.

To overcome this issue, mutual-model-based learning methods[17] have been proposed using multiple models obtained with different initial weight configurations or different portions of training data[18,19]. In this framework, each model iteratively retrains using pseudo-labels generated by other

 doi:10.20944/preprints202511.0348.v1

models rather than relying on its own predictions. This ensemble approach enhances the self-training mechanism, as each model supervises the others by providing pseudo-labels on unlabeled data, allowing for cross-supervision that helps identify and reduce errors in the pseudo-labels. Based on this methodology, we propose a dynamic mutual learning strategy that adjusts the learning weights assigned to each model through a dynamic allocation loss function, thereby improving the overall effectiveness and robustness of the learning process.

2.2. Adversarial Learning for Semi-Supervised Semantic Segmentation

Generative Adversarial Networks (GANs)[20], which comprise a generator and a discriminator, typically engage in a Min-Max game throughout the training process. This methodology has significantly contributed to the success of deep learning techniques in various image processing tasks[21,22], including semi-supervised semantic segmentation. Adversarial learning based semi-supervised semantic segmentation methods can be categorized into two distinct groups: generative and non-generative methods.

Generative methods[23,24] aim to extract and assimilate knowledge from a large corpus of unlabeled data and enhance the training dataset by generating new images. Essentially, a GAN consists of a generative component designed to mimic the distribution of the target image, thereby enabling the creation of new training examples. Simultaneously, the segmentation network acts as a discriminator, processing both real and artificially generated labels. However, since the generation method relies heavily on large volumes of unlabeled data, it can introduce significant distortion errors, leading to error accumulation.

Conversely, non-generative methods[25–28] replace the typical generative network of a classical GAN with a segmentation network. The output of this network is then fed to a discriminator that differentiates between real segmentation maps and those produced by the segmentation network. Zhang et al.[29] proposed an adaptive GAN-based segmentation algorithm aimed at improving the accuracy and generalizability of semi-supervised models. Fei et al.[30] integrated an attention mechanism into the semi-supervised learning framework for GANs, resulting in more precise predictions. Soul et al.[23] attempted to employ GANs for semi-supervised semantic segmentation, but their results were suboptimal due to significant deviations from authentic images. Luc et al.[31] introduced a training protocol for semantic segmentation using GANs that closely aligns with the framework discussed here. Inspired by above non-generative methods, we incorporated the adversarial learning to improve the evaluation of noise introduced during the pseudo-label generation stage, thereby reducing errors associated with pseudo-labels in a dynamic mutual learning context.

3. Methodology

3.1. Overview

Figure 1 illustrates our DMAS method, which comprises a two-stage training process: adversarial pre-training stage and dynamic mutual learning stage.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 November 2025 doi:10.20944/preprints202511.0348.v1

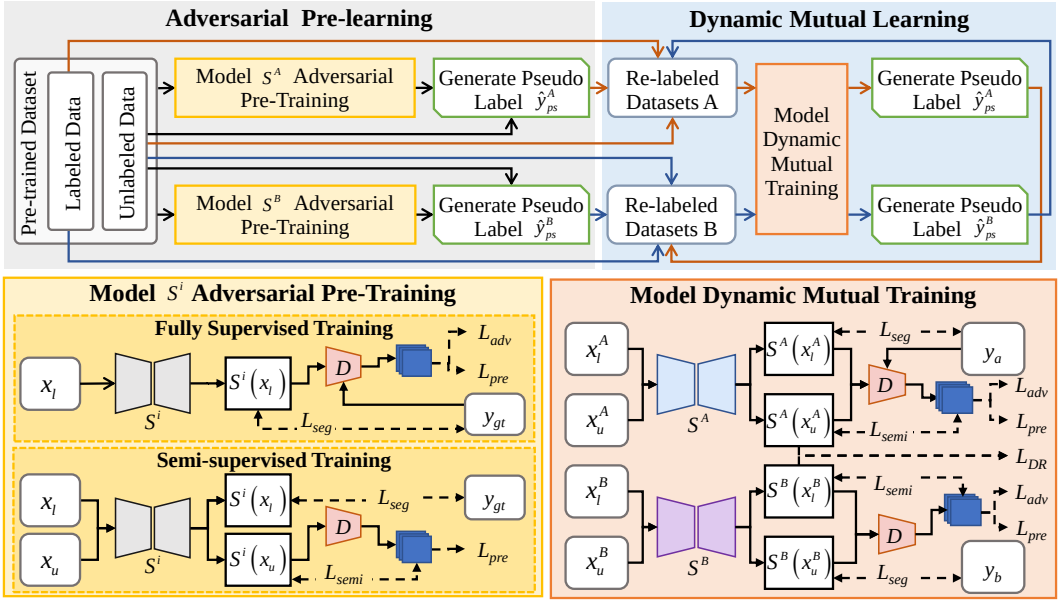

Figure 1. Overall framework of dynamic mutual adversarial segmentation(DMAS) method.The DMAS method consists of adversarial pre-training and dynamic mutual learning phases, enabling models A and B to exchange pseudo-supervision.

The adversarial pre-training stage uses a pre-training dataset divided into limited labeled data $\mathcal{I}_l$ and a larger set of unlabeled data $\mathcal{I}_u$ to develop two pre-trained segmentation networks $S^i(\cdot), i \in \{A, B\}$ and a discriminator $D(\cdot)$ capable of providing a quality evaluation, and to generate initial pseudo labels $\hat{y}^i_{ps}$. This stage includes two steps: fully supervised training and semi-supervised training. Initially, the segmentation model $S^i(\cdot)$ and the discriminator $D(\cdot)$ are fully supervised using images $x_l \in \mathcal{I}_l$ and their corresponding labels y_{gt}, within a generative adversarial training framework where $S^i(\cdot)$ acts as the generator. Then, using images $x_u \in \mathcal{I}_u$, the probability maps $D(S^i(x_u))$ from the trained discriminator serves as a supervision signal for semi-supervised training of $S^i(\cdot)$. High-probability outputs are selected as initial pseudo labels $\hat{y}^i_{ps}$. We differentiate $S_A(\cdot)$ and $S_B(\cdot)$ by integrating prior knowledge from the ImageNet dataset into one of them.

In the dynamic mutual learning stage, the pre-trained segmentation networks $S^i(\cdot), i \in \{A, B\}$ provide mutual supervision through dynamic self-training on re-labeled datasets. These datasets comprise initial labeled data $\mathcal{I}_l$, reliable pseudo-labeled data $\mathcal{I}^i_{ps}$, and remaining unlabeled data $\mathcal{I}^i_u$. Training involves using images $x^i_l \in \mathcal{I}_l \cup \mathcal{I}^i_{ps}$ and $x^i_u \in \mathcal{I}^i_u$, with labels $y_i \in \mathcal{I}_l \cup \mathcal{I}^j_{ps}$, and the probability map $D(S^i(x_u))$ as supervision signals for $S^i(\cdot)$. Pseudo labels $\hat{y}^i_{ps}$ are then generated to update the pseudo-labeled data $\mathcal{I}^j_{ps}, j \in \{A, B\}, j \neq i$ and unlabeled data $\mathcal{I}^j_u$ of the other re-labeled dataset. A dynamic re-weighting loss L_{DR} is introduced, utilizing the discrepancy between predicted confidence map and the discriminator-generated probability map $D(S^i(x_u))$ to learn from the unlabeled data.

The purpose of this stage is to iteratively enhance the segmentation network models.The details on each component of our DMAS method will be described next.

3.2. Adversarial Pre-training

The adversarial pre-training stage employs a Generative Adversarial Network (GAN) structure to enhance the quality of generated pseudo-labels, consisting of the segmentation network $S^i(\cdot)$ and the discriminator $D(\cdot)$. Subsequent pseudo-labels $\hat{y}^i_{ps}$ are obtained by calculating probabilities through the discriminator on images generated by the segmentation network and selecting the top 10% of images. In this stage, both segmentation networks $S^i(\cdot), i \in \{A, B\}$ and the discriminator $D(\cdot)$ follow the same training process, which is divided into two steps: fully-supervised training and semi-supervised training, as detailed in Algorithm 1.

 doi:10.20944/preprints202511.0348.v1

Algorithm 1 Model adversarial training.

Input: Labeled data $\mathcal{I}_l$; Unlabeled data $\mathcal{I}_u$
Output: Trained segmentation network $S(\cdot)$; Trained discriminator $D(\cdot)$
for number of fully supervised training iterations **do**
  for $x_l \in \mathcal{I}_l, y_{gt} \in \mathcal{I}_u$ **do**
    •Training segmentation network:
    argmin$(L_{seg}(y_{gt}, S(x_l)) + \lambda_{pre} L_{pre}(D(S(x_l))))$
    •Training discriminator:
    argmin$(L_{adv}(D(S(x_l)), D(y_{gt})))$
  end for
end for
for number of semi-supervised training iterations **do**
  for $x_{all} \in \mathcal{I}_l \cup \mathcal{I}_u, x_u \in \mathcal{I}_u, x_l \in \mathcal{I}_l, y_{gt} \in \mathcal{I}_l$ **do**
    •Training segmentation network:
    argmin$(L_{seg}(y_{gt}, S(x_l)) + \lambda_{pre} L_{pre}(D(S(x_{all}))) + \lambda_{semi} L_{semi}(D(S(x_u))))$
  end for
end for

3.2.1. Fully Supervised Training

In the fully-supervised training step, the objective is to train the initial segmentation networks $S^i(\cdot), i \in \{A, B\}$ and the evaluation-capable discriminator $D(\cdot)$ using only labeled data $\mathcal{I}_l$. The labeled data includes images x_l and their corresponding labels y_{gt}. We use the segmentation network $S^i(\cdot)$ to predict segmentation results $S^i(x_l)$ from images $x_l \in \mathcal{I}_l$, which contain confidence information for each class. The discriminator $D(\cdot)$ evaluates predicted results $S^i(x_l)$ with the corresponding probability maps $D(S^i(x_l))$. To maintain the independence of the segmentation network and the discriminator, an alternating training strategy is employed: in each iteration, the segmentation network is trained while keeping the discriminator parameters fixed, and then the discriminator is trained while keeping the segmentation network parameters fixed. We use a multi-class segmentation loss function $L_{seg}(y_{gt}, S(x_l))$ and a prediction loss function $\lambda_{pre} L_{pre}(D(S(x_l)))$ to optimize the segmentation network $S^i(\cdot)$ to deceive the discriminator; the adversarial loss L_{adv} is used to optimize the discriminator $D(\cdot)$ to detect subtle differences between predicted segmentation images and true labels. After fully-supervised training, the segmentation network attains initial segmentation capability, and the discriminator acquires the ability to evaluate the reliability of predicted segmentation images.

3.2.2. Semi-Supervised Training

In the semi-supervised training step, the model's performance is further enhanced using unlabeled data $\mathcal{I}_u$. The discriminator, which is proficient at distinguishing predictions from the segmentation network, generates probability maps that can act as a supervisory signal to identify areas consistent with the true label distribution. A threshold is applied to convert this map into a binary format, highlighting regions of high confidence. These regions, now serving as pseudo-labels, facilitate the self-training of the model, with the most effective iteration being determined through comparative evaluation against previous iterations. The semi-supervised training step focuses on minimizing the semi-supervised loss L_s while keeping the discriminator's parameters fixed:

$$L_s = L_{seg} + \lambda_{pre} L_{pre} + \lambda_{semi} L_{semi} \tag{1}$$

where L_{seg} and L_{pre} are the multi-class segmentation loss function and the prediction loss function, respectively, and L_{semi} represents the semi-supervised multi-class cross-entropy loss. The hyperparameters λ_{pre} and λ_{semi} are used to balance the weights of the terms. L_{seg} and L_{pre} are the same as in supervised training. The semi-supervised cross-entropy loss L_{semi} is defined as:

$$L_{semi} = -\sum_{h,w} \sum_{c \in C} I(D(S(x_u)) > T_{semi}) \cdot m \log(S(x_u)) \tag{2}$$

 doi:10.20944/preprints202511.0348.v1

where $I(\cdot)$ is the indicator function for high-probability pixel classification, while T_{semi} is the threshold that controls the sensitivity of the semi-supervised process. By binarizing the probability map $D(S(x_u))$, we can balance the credibility of generated pseudo-labels against the amount of data; specifically, a larger threshold T_{semi} can increase the credibility of the pseudo-labels, but the number of usable pseudo-labels will decrease accordingly. To enhance the utilization of unlabeled data, we set $T_{semi} = 0.2$. Since $S(x_u)$ employs one-hot encoding, we set $m^{(h,w,c)}$ to control category input. Therefore, when $c^* = \underset{c}{\text{argmax}} S(x_u)^{(h,w,c)}$, it follows that $m^{(h,w,c^*)} = 1$; otherwise, $m^{(h,w,c^*)} = 0$.

3.3. Dynamic Mutual Learning

As depicted in Figure 1, during the Dynamic Mutual Learning (DML) stage, the segmentation network parameters are progressively refined through iterative updates of the dynamic mutual learning algorithm applied to the re-labeled datasets A and B. Dataset $i \in A, B$ consists of initial labeled data $\mathcal{I}l$, re-labeled data $\mathcal{I}^i ps$, data iteratively updated from the other model $S^j(\cdot)$ where $j \in A, B$ and $j \neq i$, and its corresponding reduced unlabeled data $\mathcal{I}_u^i$. By analyzing the unlabeled images $x_u^i \in \mathcal{I}_u^i$, labeled images $x_l^i \in \mathcal{I}_l \cup \mathcal{I}_{ps}^i$, and corresponding labels $y_i \in \mathcal{I}_l \cup \mathcal{I}_{ps}^i$, the models iteratively update, leveraging differences to detect errors in automatically generated pseudo labels. Additionally, a dynamic reweighting loss L_{DR} is introduced to account for the discrepancies between the two models. Each model $S^i(\cdot)$ is trained using pseudo labels $\hat{y}_{ps}^j$ generated by the other model $S^j(\cdot)$; thus, greater pixel-level discrepancies suggest higher error likelihoods and require reduced weighting in the loss function. This minimizes the impact of those pixels or regions with significant differences on the training process. The subsequent section details the dynamic mutual iterative framework and the dynamic reweighting loss.

3.3.1. Dynamic Mutual Iterative Framework

The dynamic mutual iteration framework employs an iterative learning model that progresses from simple to complex tasks. It involves repeatedly executing a "self-training" algorithm over multiple cycles, each containing numerous low-confidence pseudo labels. Effectively utilizing these pseudo labels is crucial for enhancing the use of unlabeled datasets. In this framework, training with deterministic labeled data and highly reliable pseudo labels aligns with previous semi-supervised training, utilizing an adversarial generation method. The primary distinction lies in how pseudo labels with lower relative reliability are used. Within a specific loop, we stabilize the segmentation model $S^A(\cdot)$ and generate pseudo labels for the unlabeled image x_u^A, denoted as $S^A(x_u^A)$. These pseudo labels are then passed through a discriminator to obtain a probability map $D(S^A(x_u^A))$. By leveraging these probability maps, we can compare pseudo labels across different iterations and refine them for training segmentation model $S_B(\cdot)$ by calculating the dynamic reweighting loss L_{DR}. After appropriate weighting, the loss L_{DR} is combined with the semi-supervised training loss L_s to derive the final loss L, aiding in the training of segmentation model $S^B(\cdot)$, as shown in Equation 3.

$$L = L_s + L_{DR} \tag{3}$$

Similarly, segmentation model $S^B(\cdot)$ can be employed to train segmentation model $S^A(\cdot)$.

3.3.2. Dynamic Reweighting Loss

We introduce the dynamic reweighting loss with inter-model disagreement as reweighting weight. Assuming segmentation model $S^A(\cdot)$ is used to train segmentation model $S^B(\cdot)$, let x_u^A represent the unlabeled images from unlabeled data $\mathcal{I}_u^A$. We define its pseudo label as $\hat{y}_{ps}^{A(h,w,c)} = S^A(x_u^A)^{(h,w,c)}$, with $C^{A^{(h,w)}} = D(S^A(x_u^A))^{(h,w)}$ being the probability map. Similarly, $\hat{y}_{ps}^{B^{(h,w,c)}} = S^B(x_u^A)^{(h,w,c)}$ is the prediction during training, with $C^B(h,w) = D(S^B(x_u^A))^{(h,w)}$ as the probability map. In the loop iteration of training the segmentation model $S^A(\cdot)$, let $p^{B^{(h,w,c)}} = \hat{y}_{ps}^{B^{(h,w,c)}} \cdot C^{B^{(h,w)}}$ represent the predicted probability of class c by $S^B(\cdot)$. The dynamic reweighting loss weight $\omega_u^{(h,w)}$ is defined as:

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 November 2025 doi:10.20944/preprints202511.0348.v1

$$\omega_u^{(h,w)} = \begin{cases} p_B, \text{if } \hat{y}_{ps}^A = \hat{y}_{ps}^B \\ p_B, \text{if } \hat{y}_{ps}^A \neq \hat{y}_{ps}^B \text{ and } C^A \geq C^B \\ 0, \text{if } \hat{y}_{ps}^A \neq \hat{y}_{ps}^B \text{ and } C^A \leq C^B \end{cases} \tag{4}$$

The dynamic reweighting loss on unlabeled samples, L_{DR}, is then defined as:

$$L_{DR} = -\sum_{h,w} \omega_u^{(h,w)} \sum_{c \in C} \hat{y}_{ps}^{A(h,w,c)} \log(\hat{y}_{ps}^{B(h,w,c)}) \tag{5}$$

In the loop iteration of training the segmentation model $S^B(\cdot)$, L_{DR} is calculated using the same method, except that the results of the two segmentation models are swapped in order.

4. Experiments

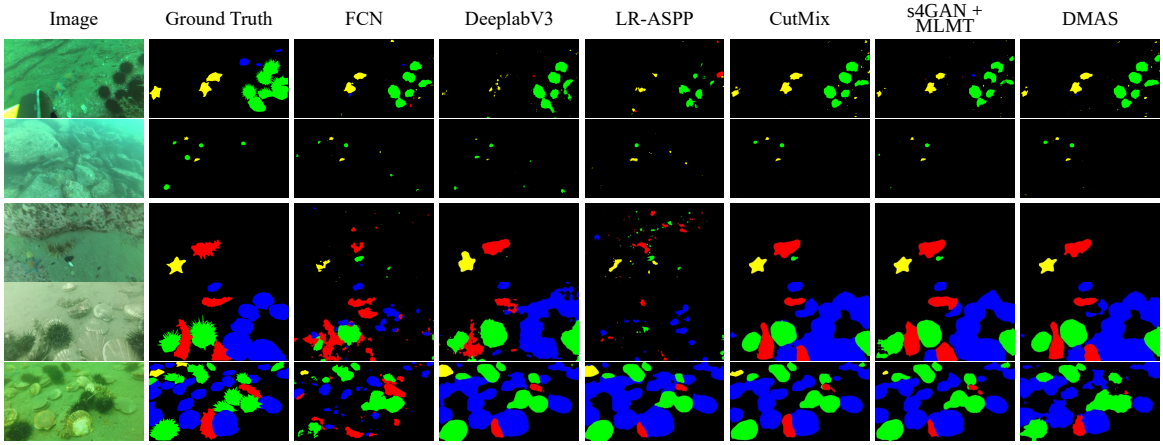

Figure 2. Visualization of comparative experimental results under 0.125 labeled ratio on UO datasets.The DMAS method demonstrates a significant advantage over the currently prevalent full-supervision semantic segmentation algorithms.

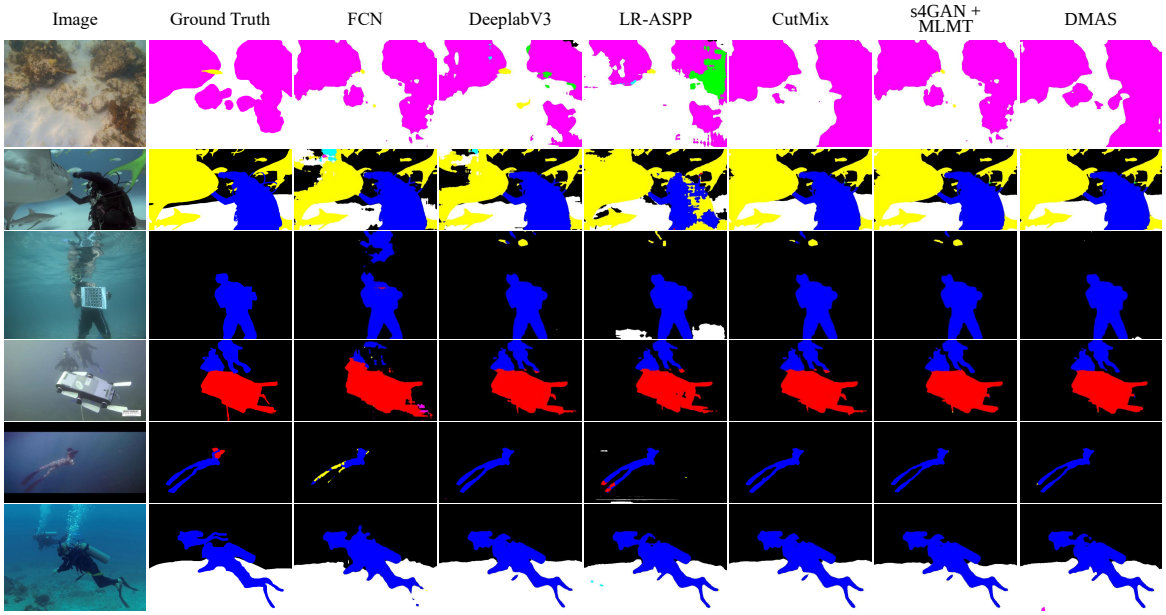

Figure 3. Visualization of comparative experimental results under 0.125 labeled ratio on SUIM datasets.DMAS excels in semi-supervised segmentation, outperforming full-supervision methods, despite minor edge delineation gaps.

 doi:10.20944/preprints202511.0348.v1

4.1. Dataset

To evaluate the effectiveness of the proposed Dynamic Mutual Adversarial Segmentation (DMAS), we conducted experiments on two publicly available underwater datasets: DUT and SUIM. The DUT dataset includes 6617 images, of which 1487 images have semantic segmentation and instance segmentation annotations, and the remaining 5130 images have object detection box annotations, categorized into four classes: starfish, holothurian, scallops, and echinus. The SUIM dataset includes 1525 images for training and validation, along with 110 test images, each with pixel-level annotations. It encompasses eight categories: fish(FS), reefs(RF), aquatic plants(PL), wrecks/ruins(WR), human divers(DR), underwater robots(RB), and seafloor(SF). Both datasets are divided into training and testing datasets in an 7:3 ratio.

In the training set, we partitioned the labeled and unlabeled data using varying labeled ratios (0.5, 0.25, 0.125, and 1). Specifically, at each random shuffle of the entire training set, the first 0.5, 0.25, and 0.125 fractions were used as labeled data. When the labeled ratio is set to 1, it denotes fully supervised learning, involving only the training and testing sets.

4.2. Implementation Details

The proposed method was implemented using PyTorch. The evaluation was conducted on a server equipped with an Intel Xeon(R) Silver 4214R CPU and an NVIDIA RTX A6000 48GB GPU. We utilized DeepLabv3 with ResNet-101[32] as the backbone. The entire model was trained for 40,000 iterations, saving the model every 200 iterations during the dynamic mutual learning. The optimal model was selected based on predictions on the test set.

For pre-training, the initial learning rate was set to 0.001, with a weight decay of 10^{-4}. The discriminator was trained using the Adam optimizer with a learning rate of 10^{-4}. For hyperparameters in the SUISS method, we set them to 0.01 and 0.001 when using labeled and unlabeled data, respectively. Random cropping, random mirroring, and other data augmentation techniques were applied to mitigate overfitting and enhance model performance. To ensure the difference between Model A and Model B, all experiments below show that Model A is an untrained network, while Model B is a pre trained network using the ImageNet dataset.

4.3. Evaluation Metrics

To objectively evaluate and analyze the performance of the DMAS method, we used the mean Pixel Accuracy (mPA) and mean Intersection over Union (mIoU) as comprehensive evaluation metrics for segmentation effectiveness. True Positive (TP) represents the number of correctly predicted samples, False Negative (FN) represents the number of falsely predicted samples, False Positive (FP) represents the number of samples falsely identified as positive, and True Negative (TN) represents the number of correctly identified negative samples. Pixel Accuracy (PA) is defined as:

$$PA = \frac{TP + TN}{FN + FP + TP + TN} \tag{6}$$

IoU indicates the intersection over the union between the predicted and the actual value for each category. mIoU is the average IoU across all categories, and its calculation is shown:

$$mIoU = \frac{1}{k+1} \sum_{i=0}^{k} IoU = \frac{1}{k+1} \sum_{i=0}^{k} \frac{TP}{FN + FP + TP} \tag{7}$$

 doi:10.20944/preprints202511.0348.v1

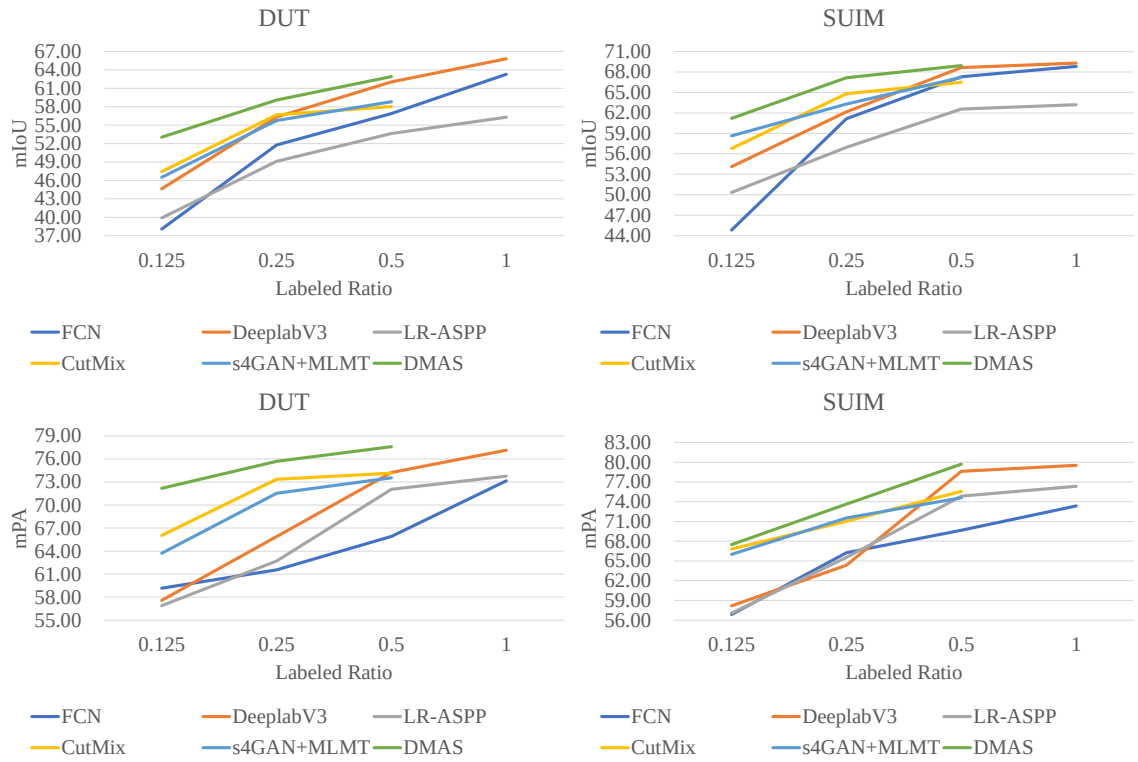

Figure 4. Comparative experimental results under different labeled ratios on DUT dataset and SUIM dataset.DMAS excels in semantic segmentation with 0.125 labeled ratio data, proving its robustness and data efficiency.

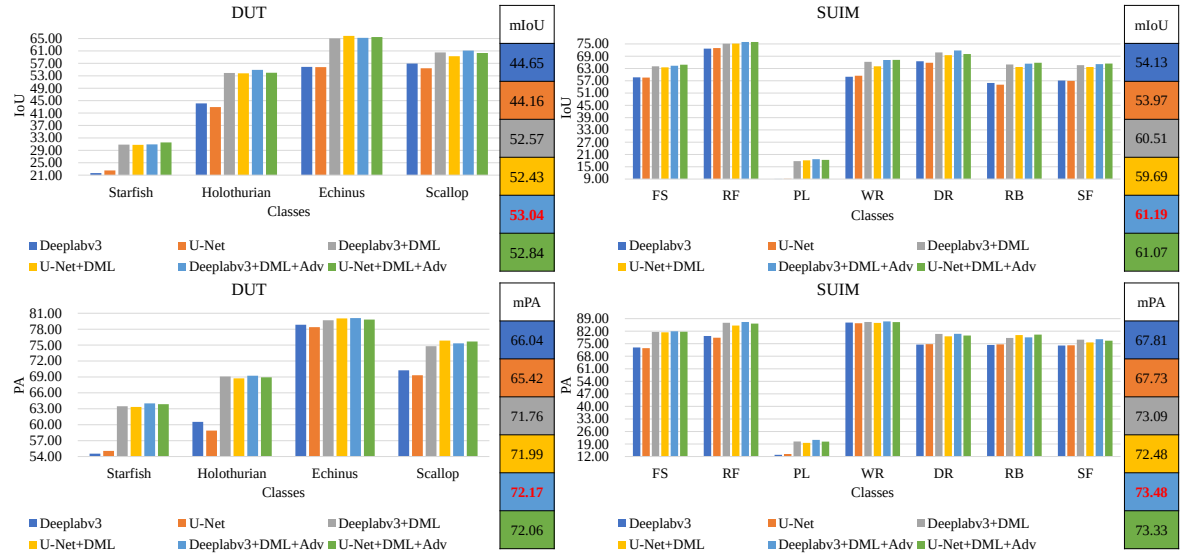

Figure 5. Quantitative comparison of ablation experiment results PA under 0.125 labeled ratio on SUIM dataset and UO dataset.Enhanced segmentation via dynamic mutual learning and adversarial strategy, achieving mPA exceeds 0.7.

4.4. Performance Comparison

We compared the DMAS method on the DUT and SUIM datasets with several widely used fully supervised semantic segmentation algorithms, including:

a) Supervised learning with fine tuning: FCN[33], DeeplabV3[23] and LR-ASPP[34] methods.

b) Semi-supervised learning: CutMix augmentation[35] and hybrid method s4GAN+MLMT[36].

 doi:10.20944/preprints202511.0348.v1

All the above algorithms were implemented on the PyTorch platform, utilizing the same NVIDIA RTX A6000 48GB GPU, and were evaluated using the same pre-training and parameters on both the DUT and SUIM datasets.

4.4.1. Quantitative Analysis

Figure 4 display the differential outcomes of semantic segmentation experiments conducted at varying labeled ratios on the DUT and SUIM datasets, respectively. Analyzing these tables reveals that our proposed method, DMAS, exhibits superior performance compared to existing supervised and semi-supervised learning approachs, even with limited data availability in the DUT and SUIM datasets. Notably, DMAS maintains a consistently high performance regardless of reduced training sample sizes. For instance, at a labeled ratio of 0.125, DMAS achieves commendable mIoU and mPA metrics, comparable to those obtained by LR-ASPP in a fully supervised training scenario. These results highlight the robustness and effectiveness of DMAS in handling scenarios with sparse labeled data. The data clearly support the hypothesis that the semi-supervised DMAS algorithm provides a significant advantage in semantic segmentation tasks, particularly when labeled ratios are limited, thus establishing its superiority in data-efficient learning models.

4.4.2. Qualitative Analysis

To provide a more tangible evaluation of DMAS's performance, we conducted training on the DUT and SUIM datasets with an annotation ratio of 0.125. The comparative results are illustrated in Figures 2 and 3. It is apparent that, even with a limited number of annotated samples, the DMAS method significantly outperforms other semantic segmentation methods based on full supervision or semi-supervision, particularly in terms of pixel classification accuracy and object boundary delineation. Specifically, as shown in the fifth row of Figure 2 and the third column of Figure 3, when various methods segment the same image, DMAS demonstrates superior performance in contour edge articulation, pixel classification over larger areas, and the accuracy of classification for smaller objects, surpassing other methods. This reflects DMAS's efficacy in leveraging unlabeled data to enhance the model's segmentation capability, thereby achieving successful semi-supervised learning. However, there remains a noticeable gap between DMAS's predictions and the actual labels, as illustrated in the first row of Figure 2 and the thirteenth row of Figure 3. The precision of edge segmentation and regional delineation for targets is still suboptimal, which may be due to inherent performance limitations of the segmentation network or an insufficiency of training data. Overall, in terms of achieving more accurate predictions for underwater image at a reduced cost, the DMAS method demonstrates a significant advantage over the currently prevalent full-supervision semantic segmentation algorithms.

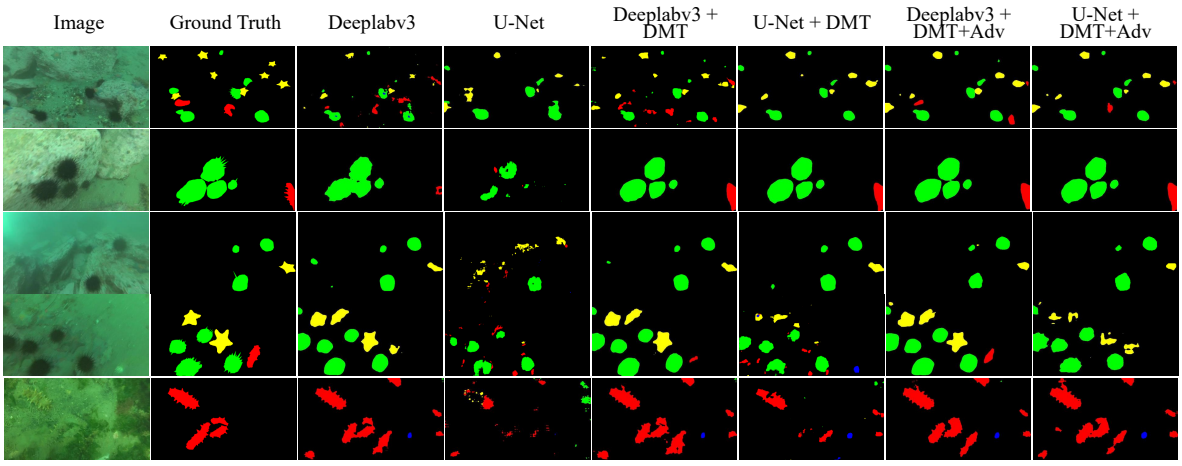

Figure 6. Visualization of ablation experimental results under 0.125 labeled ratio on UO datasets.models trained using the DMAS method can segment the majority of underwater target areas more proficiently.

 doi:10.20944/preprints202511.0348.v1

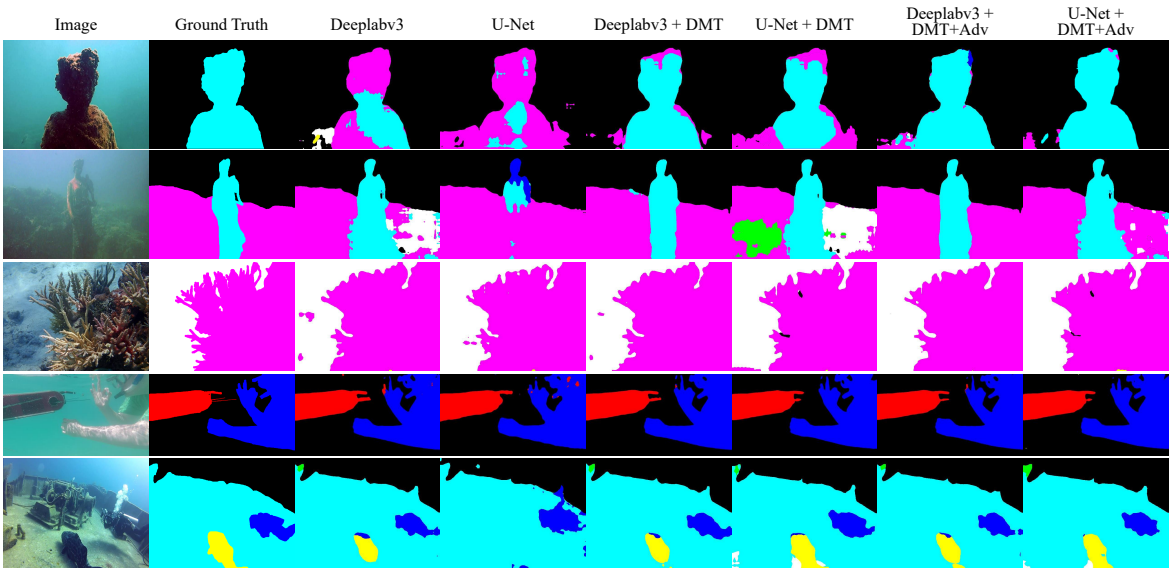

Figure 7. Visualization of ablation experimental results under 0.125 labeled ratio on SUIM datasets.Visualization of ablation experimental results under 0.125 labeled ratio on SUIM datasets.DMAS models excel in segmenting low-contrast underwater targets and peripheral details, reducing classification errors. However, concealed targets still present challenges.

4.5. Ablation Study

To further assess the rationality of the framework design choice in the DMAS method, we conducted ablation experiments on two underwater datasets. Initially, the baseline network consists of a fully supervised trained model within the DMAS method's segmentation network, including Deeplabv3 and U-Net. Subsequently, this baseline network is enhanced with dynamic mutual learning. In contrast, the DMAS method not only incorporates the baseline network but also introduces model adversarial training and a dynamic mutual learning learning strategy. Under a labeling rate of 0.125, these experiments employed the same pre-trained dataset, trained on the training dataset, and subsequently evaluated on the test dataset. The quantitative analysis of the test sets for both datasets is illustrated in Figure 5. Training with the dynamic mutual learning method within the same experimental setup has significantly enhanced the model's segmentation precision compared to solely using the baseline network. Concurrently, the inclusion of an adversarial training strategy has improved the algorithm's overall performance. For discernible underwater targets, the mean Pixel Accuracy (mPA) exceeds 0.7, and for well-segmented instances, it can reach up to 0.8. This indicates that the dynamic mutual learning strategy can effectively utilize unlabeled data to extract associative information, thereby optimizing the segmentation model.

As illustrated by the visualization results in Figures 6 and 7, models trained using the DMAS method can segment the majority of underwater target areas more proficiently. When faced with low-contrast and indistinct contour targets (as shown in the third row of Figure 6 and the second row of Figure 7), the models exhibit minimal classification errors and perform superior segmentation of small targets (as shown in the fourth row of Figure 6) and certain peripheral details. In the case of semantic segmentation of concealed targets (as shown in rows one and five of Figure 6), numerous errors still persist, yet there is a relative enhancement in classification accuracy compared to the other two methods.

4.6. Discussion

The increasing demand for advanced underwater detection and monitoring technologies has made underwater image semantic segmentation a prominent research area. Researchers are increasingly turning to convolutional neural networks (CNNs) and their variants. These models enhance the accuracy and robustness of segmentation. These models automatically learn hierarchical features,

 doi:10.20944/preprints202511.0348.v1

making them well-suited for addressing the complex patterns and textures characteristic of underwater scenes.

Despite these advancements, the field of underwater image semantic segmentation continues to face numerous challenges: 1) The lack of large annotated datasets limits the training and validation of deep learning algorithms. 2) Factors such as low visibility, poor contrast, and the presence of various underwater elements often contribute to significant annotation noise, thereby obscuring the segmentation results.

To tackle the challenges posed by limited noise annotations, many studies have proposed self-training methods, wherein models iteratively label and retrain on their predicted results. This method has been applied to underwater image segmentation. Especially when combined with confidence thresholds, this method can gradually enhance the model's performance on noisy data. However, a notable limitation of self-training methodes is the lack of mechanisms to detect and correct errors during the training process. Furthermore, the choice of confidence threshold directly impacts the effective utilization of low-confidence pseudo-labels.

To address the issue of underwater image semantic segmentation with limited noisy annotations, we have developed more sophisticated solutions. By leveraging collaborative platforms and crowd-sourcing techniques, we can aggregate multiple annotations, which helps reduce noise levels in the training dataset. Aggregated data validated through multiple sources can provide higher quality annotations. We enable dynamic mutual learning between two distinctly different models. Each model iteratively retrains on unannotated data, using pseudo-labels generated by the other model. This process allows the models to guide each other, enhancing their noise correction capabilities. Simultaneously, in selecting the confidence threshold for pseudo-label selection, we no longer rely on traditional confidence levels. Instead, inspired by generative adversarial concepts, we train a discriminator to focus on the performance of pseudo-labels, thereby making their selection more adaptive.

Our method significantly reduces noise levels in underwater image semantic segmentation, but there are still limitations that need to be addressed. Although dynamic mutual learning theoretically improves noise correction in models, its practical application can cause instability during training due to interdependencies among models, which affects convergence speed and overall performance. This is particularly evident in model selection, which consequently affects the efficacy of dynamic mutual learning. To tackle these challenges, we intend to integrate more advanced semantic segmentation network architectures, including multi-scale feature extraction and multi-task learning, to enhance the model's ability to handle complex underwater environments. Furthermore, we plan to explore multimodal learning methods to improve the model's robustness and generalization by integrating diverse sensor data, including vision and sonar.

5. Conclusions

This paper introduces a novel semi-supervised method for underwater image semantic segmentation, named Dynamic Mutual Adversarial Segmentation (DMAS). The method employs a two-stage framework: adversarial pre-training and dynamic mutual learning. In the adversarial pre-training stage, the segmentation model is initially trained using labeled data to generate pseudo-labels. Concurrently, a convolutional discriminator is trained on both the segmentation maps derived from labeled and unlabeled data. This dual training process enables the network to differentiate between actual and predicted labels by producing a confidence map, which aids in qualitatively assessing the segmentation accuracy.During the dynamic mutual learning stage, multiple models with different prior knowledge are trained. By analyzing the divergence between these models, errors in the pseudo-labels can be detected. The training process incorporates a dynamically reweighted loss function, which adjusts weights based on the differences between the models. This method reduces the impact of discrepancies that suggest potential errors, thereby improving the precision of the training process.Experimental results on the DUT and SUIM datasets indicate that DMAS significantly enhances semantic segmentation performance, even with a limited amount of labeled data.

 doi:10.20944/preprints202511.0348.v1

References

1. B. P. McNelly, "Advances in autonomous underwater vehicle technologies for enhanced harbor protection," Ph.D. dissertation, Johns Hopkins University, 2023.
2. Z. Li, S. Liang, M. Guo, H. Zhang, H. Wang, Z. Li, and H. Li, "Adrc-based underwater navigation control and parameter tuning of an amphibious multirotor vehicle," *IEEE Journal of Oceanic Engineering*, 2024.
3. X. Yang, Z. Song, I. King, and Z. Xu, "A survey on deep semi-supervised learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 9, pp. 8934–8954, 2022.
4. H. Zeng, Z. Liu, and H. Cai, "Research on the application of deep learning in computer network information security," in *Journal of Physics: Conference Series*, vol. 1650, no. 3. IOP Publishing, 2020, p. 032117.
5. M. Zhang, Y. Zhou, J. Zhao, Y. Man, B. Liu, and R. Yao, "A survey of semi-and weakly supervised semantic segmentation of images," *Artificial Intelligence Review*, vol. 53, pp. 4259–4288, 2020.
6. Q. Li, X.-M. Wu, H. Liu, X. Zhang, and Z. Guan, "Label efficient semi-supervised learning via graph filtering," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 9582–9591.
7. Y. Wei, H. Xiao, H. Shi, Z. Jie, J. Feng, and T. S. Huang, "Revisiting dilated convolution: A simple approach for weakly-and semi-supervised semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7268–7277.
8. Y. Ouali, C. Hudelot, and M. Tami, "Semi-supervised semantic segmentation with cross-consistency training," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 12 674–12 684.
9. D.-H. Lee *et al.*, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *Workshop on challenges in representation learning, ICML*, vol. 3, no. 2. Atlanta, 2013, p. 896.
10. R. Li, S. Li, C. He, Y. Zhang, X. Jia, and L. Zhang, "Class-balanced pixel-level self-labeling for domain adaptive semantic segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 11 593–11 603.
11. I. Triguero, S. García, and F. Herrera, "Self-labeled techniques for semi-supervised learning: taxonomy, software and empirical study," *Knowledge and Information systems*, vol. 42, pp. 245–284, 2015.
12. S.-S. Learning, "Semi-supervised learning," *CSZ2006. html*, vol. 5, p. 2, 2006.
13. L. Yang, W. Zhuo, L. Qi, Y. Shi, and Y. Gao, "St++: Make self-training work better for semi-supervised semantic segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 4268–4277.
14. E. W. Teh, T. DeVries, B. Duke, R. Jiang, P. Aarabi, and G. W. Taylor, "The gist and rist of iterative self-training for semi-supervised segmentation," in *2022 19th Conference on Robots and Vision (CRV)*. IEEE, 2022, pp. 58–66.
15. H. Li and H. Zheng, "A residual correction approach for semi-supervised semantic segmentation," in *Pattern Recognition and Computer Vision: 4th Chinese Conference, PRCV 2021, Beijing, China, October 29–November 1, 2021, Proceedings, Part IV 4*. Springer, 2021, pp. 90–102.
16. J. Yuan, Y. Liu, C. Shen, Z. Wang, and H. Li, "A simple baseline for semi-supervised semantic segmentation with strong data augmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 8229–8238.
17. Y. Zhang, T. Xiang, T. M. Hospedales, and H. Lu, "Deep mutual learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4320–4328.
18. Z. Feng, Q. Zhou, Q. Gu, X. Tan, G. Cheng, X. Lu, J. Shi, and L. Ma, "Dmt: Dynamic mutual training for semi-supervised learning," *Pattern Recognition*, vol. 130, p. 108777, 2022.
19. Y. Zhou, R. Jiao, D. Wang, J. Mu, and J. Li, "Catastrophic forgetting problem in semi-supervised semantic segmentation," *IEEE Access*, vol. 10, pp. 48 855–48 864, 2022.
20. A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, "Generative adversarial networks: An overview," *IEEE signal processing magazine*, vol. 35, no. 1, pp. 53–65, 2018.
21. J. Gui, Z. Sun, Y. Wen, D. Tao, and J. Ye, "A review on generative adversarial networks: Algorithms, theory, and applications," *IEEE transactions on knowledge and data engineering*, vol. 35, no. 4, pp. 3313–3332, 2021.
22. Z. Wang, Q. She, and T. E. Ward, "Generative adversarial networks in computer vision: A survey and taxonomy," *ACM Computing Surveys (CSUR)*, vol. 54, no. 2, pp. 1–38, 2021.
23. N. Souly, C. Spampinato, and M. Shah, "Semi supervised semantic segmentation using generative adversarial network," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5688–5696.
24. D. Li, J. Yang, K. Kreis, A. Torralba, and S. Fidler, "Semantic segmentation with generative models: Semi-supervised learning and strong out-of-domain generalization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 8300–8311.

 doi:10.20944/preprints202511.0348.v1

25. G. Jin, C. Liu, and X. Chen, "Adversarial network integrating dual attention and sparse representation for semi-supervised semantic segmentation," *Information Processing & Management*, vol. 58, no. 5, p. 102680, 2021.
26. D. Xu and Z. Wang, "Semi-supervised semantic segmentation using an improved generative adversarial network," *Journal of Intelligent & Fuzzy Systems*, vol. 40, no. 5, pp. 9709–9719, 2021.
27. R. Mendel, L. A. De Souza, D. Rauber, J. P. Papa, and C. Palm, "Semi-supervised segmentation based on error-correcting supervision," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIX 16*. Springer, 2020, pp. 141–157.
28. Z. Ke, D. Qiu, K. Li, Q. Yan, and R. W. Lau, "Guided collaborative training for pixel-wise semi-supervised learning," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16*. Springer, 2020, pp. 429–445.
29. W.-C. Hung, Y.-H. Tsai, Y.-T. Liou, Y.-Y. Lin, and M.-H. Yang, "Adversarial learning for semi-supervised semantic segmentation," *arXiv preprint arXiv:1802.07934*, 2018.
30. J. Zhang, Z. Li, C. Zhang, and H. Ma, "Stable self-attention adversarial learning for semi-supervised semantic image segmentation," *Journal of Visual Communication and Image Representation*, vol. 78, p. 103170, 2021.
31. P. Luc, C. Couprie, S. Chintala, and J. Verbeek, "Semantic segmentation using adversarial networks," *arXiv preprint arXiv:1611.08408*, 2016.
32. S. C. Yurtkulu, Y. H. Şahin, and G. Unal, "Semantic segmentation with extended deeplabv3 architecture," in *2019 27th Signal Processing and Communications Applications Conference (SIU)*. IEEE, 2019, pp. 1–4.
33. J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.
34. A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan *et al.*, "Searching for mobilenetv3," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 1314–1324.
35. G. French, S. Laine, T. Aila, M. Mackiewicz, and G. Finlayson, "Semi-supervised semantic segmentation needs strong, varied perturbations," *arXiv preprint arXiv:1906.01916*, 2019.
36. S. Mittal, M. Tatarchenko, and T. Brox, "Semi-supervised semantic segmentation with high-and low-level consistency," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 4, pp. 1369–1379, 2019.