Dual Adaptive Augmentation with Difficulty-Aware Guidance for Semi-Supervised Semantic Segmentation

Wen Yan[1], Jinghua Zhu[1*], Heran Xi[1]
[1]School of Computer Science and Technology, Heilongjiang University, Harbin, Heilongjiang, China
Email: 2232012@s.hlju.edu.cn, zhujinghua@hlju.edu.cn, xiheran@hlju.edu.cn

***Abstract*—Semi-supervised semantic segmentation has made significant advances, yet current methods often overlook variations in sample difficulty and rely on limited data augmentation strategies, which constrains both accuracy and the exploration of diverse perturbation spaces. This paper proposes a Dual Adaptive Augmentation framework. First, we introduce a quantitative hardness evaluation mechanism based on the Intersection over Union (IoU) of segmentation predictions, which assesses the learning difficulty of each unlabeled sample. Guided by the derived hardness coefficient, we dynamically adjust the mixing ratio of weak and strong augmentations, aligning perturbation intensity with the model's evolving learning state. Second, we expand the perturbation space by applying two independent strong augmentation strategies to unlabeled images, thereby improving generalization. Finally, we introduce additional random perturbations, such as Dropout2D, between the encoder and decoder layers to enhance feature-level diversity and robustness. Extensive experiments demonstrate that our model consistently outperforms existing methods across different data partition settings.**

***Index Terms*—Semi-Supervised Learning, Semantic Segmentation, Data Augmentation**

I. Introduction

The goal of semi-supervised semantic segmentation (SSS) [1], [2] is to train models with a limited number of labeled images while effectively leveraging large amounts of unlabeled data. This approach aims to reduce dependence on expensive manual annotations, lower training costs, and improve overall efficiency.

The mainstream approach in semi-supervised semantic segmentation (SSS) is the Consistency Regularization Framework [3]–[6], Based on this framework, studies have shown [4], [5] that introducing strong perturbations (e.g., Mixup [7]) during SSS training can significantly improve model performance. For example, FixMatch [8] applies supervision to strongly augmented unlabeled images, enabling the model to extract richer visual information. Its effectiveness lies in generating high-quality pseudo-labels on weakly augmented data to guide the model's gradual adaptation to stronger perturbations. However, FixMatch fails to consider the varying sensitivity of different samples to perturbations and relies solely on a single augmentation path. Although iMAS [9] introduces some improvements in augmentation strategies, it still employs a single-path design, limiting the diversity of augmentation and the expressive capacity of the perturbation space. These limitations highlight the need for a unified framework that accounts for sample-specific differences and integrates both image-level and feature-level perturbations to form a multi-level augmentation mechanism.

To this end, we propose a dual adaptive augmentation model for semi-supervised semantic segmentation. First, we introduce a class-weighted symmetric IoU (sIoU) metric based on teacher–student predictions to dynamically assess sample difficulty, capturing both the model's generalization ability and the evolving difficulty during training. Next, we adopt a dual-pipeline augmentation strategy, applying two randomly selected strong augmentations to unlabeled data. The mixing ratio between strong and weak augmentations is dynamically adjusted according to sample difficulty, ensuring that perturbation strength matches the model's learning state. Loss weights are also adapted, allowing the model to prioritize high-confidence samples and gradually address harder ones, while dual-path perturbations help mitigate overfitting. Finally, we introduce feature-level perturbations between the encoder and decoder, complementing image-level augmentations to establish dual-level consistency constraints in both pixel and feature spaces.The key contributions of this work are:

- We propose a dual-adaptive augmentation model for semi-supervised semantic segmentation that incorporates instance hardness into loss and perturbation to adapt dynamically during training.
- We introduce a prediction-consistency-based hardness quantification method that assesses sample difficulty using the IoU of model predictions, overcoming the limitations of uniform augmentation by aligning perturbation intensity with actual learning progress.
- We design a dual-level perturbation mechanism that combines two strong image augmentations with feature-level random perturbations (e.g., Dropout2D), thereby expanding the diversity of the perturbation space and improving model robustness in an end-to-end manner.

II. Related Work

A. Semi-supervised Semantic Segmentation

Semi-supervised learning is focused on how to effectively use unlabeled data to improve the model learning ability

*Corresponding author

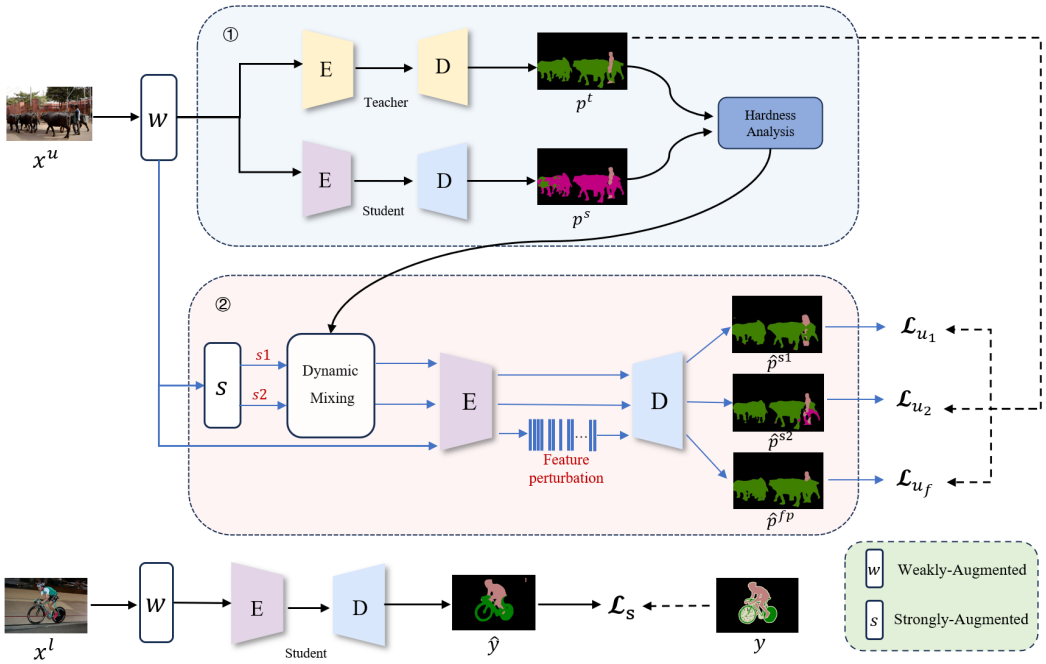

Fig. 1. Overview of the proposed segmentation framework.

[10]–[12]. In recent years, Consistency Regularization (CR) [6], [13]–[15] has become an important technique in semi-supervised learning, allowing models to train simultaneously on labeled and unlabeled data, thus improving generalization ability. To enhance the performance of semi-supervised semantic segmentation, many studies have explored methods to strengthen supervision signals. For example, some methods introduce extra losses terms [3] to optimize learning objectives, or employ stronger data augmentation strategies [16] to increase data diversity. In addition, advanced contrastive learning methods [17]–[19] have been applied to uncover latent signals in unlabeled data. CPS [3] improves the segmentation ability of the model through mutual supervision of dual models. On the other hand, U²PL [14], inspired by contrastive learning, treats uncertain pixels as negative samples and expands the perturbation space to align with the constraints of existing methods. However, these methods generally face two issues: significant variations in the learning difficulty between unlabeled data and a limited perturbation space. To address these challenges, we propose an adaptive augmentation strategy and expand the perturbation space to overcome these limitations.

B. Data Augmentation

Currently, there are three main directions to enhance SSS performance, including augmentations, more supervision, and pseudo-rectifying. Most of the existing research uses various strong data augmentations to perturb unlabeled data. One such augmentation method is AutoAugment [20], which automatically searches for the optimal augmentation strategy to improve the model's training performance. AutoAugment uses reinforcement learning to choose the most appropriate augmentation operations and strengths for specific tasks. However, the goal of AutoAugment's augmentation is to find the best augmentation operations and strengths. In contrast, in SSS, the purpose of applying data augmentation is to generate different inputs without specific goals or search spaces. Additionally, directly applying these augmentations may over-distort unlabeled data and damage the data distribution, thus degrading performance. Therefore, in this work, we propose a simple and effective data augmentation method to improve SSS performance. We use a highly random design to simplify the standard RandomAug [21], avoiding complex strategy search and generating different augmentation operations for each mini-batch during training, thus improving model robustness and generalization ability.

III. Method

A. Overview

The core objective of semi-supervised semantic segmentation is to leverage a limited set of labeled data $D_s = \{(x_i^l, y_i^l)\}$ to effectively learn from a large number of unlabeled images $D_u = \{x_i^u\}$. Typically, $|D_s| \ll |D_u|$. During each iteration, we randomly sample a batch from the labeled dataset: $D_s = \{(x_i^l, y_i^l)\}_{i=1}^{D_s}$, as well as another batch from the unlabeled dataset: $D_u = \{(x_i^u)\}_{i=1}^{D_u}$. The general training loss function is defined as follows:

$$\mathcal{L} = \mathcal{L}_s + \lambda_u \mathcal{L}_u. \tag{1}$$

where λ_u is a hyperparameter used to balance the weight between the supervised loss $\mathcal{L}_x$ on labeled samples and the unsupervised loss $\mathcal{L}_u$ on unlabeled samples.

B. Hardness assessment

We compute a symmetric class-weighted IoU between the predictions of the teacher and student models to dynamically assess sample difficulty. This method simultaneously mitigates

class imbalance and enhances the model's adaptability to different samples during training. As shown in fig. 1, for each weakly augmented unlabeled sample, we obtain the segmentation predictions p_t^i and p_s^i from the teacher and student models, respectively.

$$p_i^s = f_{\theta_s}(\mathcal{A}_w(x_i^u)), \quad \rho_i^s = \frac{1}{H \times W} \sum_{j=1}^{H \times W} \mathbb{1}(\max(p_i^s(j)) \geq \tau). \tag{2}$$

$$p_i^t = f_{\theta_t}(\mathcal{A}_w(x_i^u)), \quad \rho_i^t = \frac{1}{H \times W} \sum_{j=1}^{H \times W} \mathbb{1}(\max(p_i^t(j)) \geq \tau). \tag{3}$$

where ρ_i^s and ρ_i^t represent the high-confidence ratios on p_i^s and p_i^t, respectively. $\mathbb{1}$ is an indicator function used to compute the proportion of pixels that satisfy the given condition. Next, we define the class-weighted IoU, denoted as wIoU, to measure the prediction consistency between the student and teacher models. This metric applies weighting for each class:

$$\text{wIoU}(z_1, z_2) = \sum_c \frac{|z_1(c)|}{H \times W} \text{IoU}(z_1(c), z_2(c)). \tag{4}$$

where $\mathbf{z_1}$ and $\mathbf{z_2}$ represent the segmentation predictions of the student and teacher models, respectively. The variable $\mathbf{c}$ denotes the class index, and $|\mathbf{z_1}(c)|$ represents the number of pixels belonging to class $\mathbf{c}$ in the segmentation result. To ensure the symmetry of the evaluation, we compute the hardness metric γ_i as:

$$\gamma_i = 1 - \left[\frac{\rho_i^s}{2} \text{wIoU}(p_i^s, p_i^t) + \frac{\rho_i^t}{2} \text{wIoU}(p_i^t, p_i^s) \right]. \tag{5}$$

where the coefficient $1/2$ ensures that γ_i remains within the range of $[0, 1]$. A higher γ_i value indicates greater difficulty and a stronger need for perturbation, whereas a lower value suggests a relatively easier-to-process sample.

C. Dual Adaptive Augmentation

To address the limitation of the perturbation space, we generate two independent perturbation streams (x^{s1}, x^{s2}) from the strong perturbation pool $\mathcal{A}_s$, both based on x^w. Since $\mathcal{A}_s$ is predefined but inherently stochastic, x^{s1} and x^{s2} are not identical. Considering that strong augmentations may introduce distributional shifts and weaken fine-grained feature representations, especially in the early stages of training, we adjust the augmentation strength of unlabeled samples by mixing the outputs of strong and weak augmentations according to the instance difficulty computed in subsection B of Method.

$$\mathcal{A}_{s1}(x_i^u) \leftarrow (1 - \gamma_i)\mathcal{A}_{s1}(x_i^u) + \gamma_i \mathcal{A}_w(x_i^u), \tag{6}$$

$$\mathcal{A}_{s2}(x_i^u) \leftarrow (1 - \gamma_i)\mathcal{A}_{s2}(x_i^u) + \gamma_i \mathcal{A}_w(x_i^u). \tag{7}$$

Through this approach, challenging samples with higher r values are constrained within a reasonable perturbation range to avoid excessive difficulty from potential out-of-distribution cases, while simpler samples with lower r values can leverage strong augmentations to further enhance learning.

D. Feature Perturbation

We construct a broader perturbation space by applying additional perturbations to the intermediate features of the student model during the weakly perturbed feature extraction phase of x^w. We adopt independent data streams to separately handle image-level and feature-level perturbations, enabling the student model to learn representations at different levels and achieve consistency. Formally, the student model f_{θ_s} can be decomposed into an encoder E_s and a decoder D_s. We obtain p^{fp} from the perturbation stream:

$$e^w = E_s(x^w), \tag{8}$$

$$p^{fp} = D_s(\mathcal{P}(e^w)). \tag{9}$$

In summary, we design three parallel data streams, as shown in Figure 1. The baseline stream provides reference predictions from the teacher model, the image-level augmentation stream enhances the model's adaptability to diverse inputs through dual strong perturbations, and the feature-level perturbation stream introduces intermediate feature perturbations in the student model to achieve multi-level consistency learning. Finally, an unsupervised loss function is applied to enforce prediction consistency across the three streams.

$$\mathcal{L}_u = \frac{1}{|\mathcal{B}_u|} \sum_{i=1}^{|\mathcal{B}_u|} \frac{1 - \gamma_i}{2H \times W} \sum_{j=1}^{H \times W} \mathbb{1}(\max(p_i^t(j)) \geq \tau) \Big(\lambda_p (CE(p_i^t(j), p^{fp})) + \lambda_s \big(CE(f_{\theta_s}(\mathcal{A}_{s1}(x_i^u)), p_i^t(j)) + CE(f_{\theta_s}(\mathcal{A}_{s2}(x_i^u)), p_i^t(j)) \big) \Big). \tag{10}$$

Where λ_p and λ_s represent the weights for the feature-level and image-level loss, respectively.

IV. Experiments

A. Implementation Details

To systematically verify the effectiveness of the proposed method, this study evaluates the segmentation performance of all runs using the mean Intersection-over-Union (mIoU) metric on two authoritative benchmarks in the field of semi-supervised semantic segmentation: PASCAL VOC 2012 [22] and Cityscapes [23]. The model adopts ResNet-101 [24], pretrained on ImageNet [25], as the encoder, and DeepLabv3+ [26] as the decoder. During training, the student model is optimized using SGD with a momentum of 0.9, an initial learning rate of 0.01, and multiple learning rate decay strategies. The model is trained for 80 epochs on the VOC 2012 dataset and 240 epochs on the Cityscapes dataset. The input data is randomly cropped to 513×513 (VOC 2012) and 769×769 (Cityscapes) resolutions to match the network input. Specifically, PASCAL VOC 2012 includes 21 semantic categories, with a base training set of 1,464 images and a validation set of 1,449 images. Furthermore, the SBD dataset [27] is incorporated to expand the training set with 9,118 weakly labeled images. Cityscapes focuses on urban scene understanding and provides 19 finely labeled categories, with a training set containing 2,975 high-resolution images (2048×1024) and a validation set of 500 precisely annotated samples.

TABLE I
PERFORMANCE COMPARISON ON THE PASCAL VOC 2012 VALIDATION SET USING THE RESNET-101 BACKBONE NETWORK.

Method	1/8 (183)	1/4 (366)	1/2 (732)	Full (1464)
PseudoSeg [15]	64.3	67.2	70.5	71.1
PC^2Seg [28]	65.2	68.0	71.3	72.2
ST++ [16]	68.8	72.1	74.2	76.5
U^2PL [14]	68.2	71.5	73.6	77.1
CPS [3]	65.8	69.7	73.3	-
PSMT [29]	68.4	74.0	75.1	77.3
S4MC [30]	72.5	75.7	77.8	78.9
iMAS [9]	72.2	75.5	77.4	79.0
ours	**73.8**	**76.6**	**78.2**	**79.6**

TABLE II
PERFORMANCE COMPARISON RESULTS ON THE PASCAL VOC 2012 DATASET USING THE HYBRID TRAINING SET.

Method	ResNet-50			ResNet-101		
	1/16	1/8	1/4	1/16	1/8	1/4
CCT [13]	63.5	68.9	71.3	66.4	71.2	74.1
GCT [31]	62.5	68.7	71.2	67.6	71.4	73.5
ST++ [16]	70.1	72.3	73.2	72.6	74.2	74.3
CPS [3]	69.9	72.0	73.9	72.5	74.4	75.1
PSMT [29]	70.0	73.6	74.4	73.2	76.5	76.7
DDFP [32]	-	-	-	76.0	76.0	77.2
S4MC [30]	-	-	-	75.5	76.0	76.9
iMAS [9]	72.4	74.3	75.0	75.0	75.8	77.0
ours	**73.5**	**75.1**	**75.3**	**76.2**	**76.7**	**77.6**

B. Comparison with State-of-the-Art Methods

As shown in table I and table II, our model demonstrates superior performance on both the standard PASCAL VOC 2012 dataset and the Hybrid VOC dataset. In particular, under low-label conditions, our model exhibits stronger generalization capability and higher data efficiency. Compared with SOTA methods such as iMAS, our model achieves better or comparable segmentation accuracy with fewer labeled samples, fully highlighting the advantages of our approach.

As shown in table III, on the Cityscapes dataset, our model, based on the ResNet-50 backbone, achieves the best performance across different partitioning settings (1/16, 1/8, 1/4), consistently outperforming existing SOTA methods and demonstrating excellent generalizability and robustness.

TABLE III
PERFORMANCE EVALUATION IS CONDUCTED ON THE CITYSCAPES VALIDATION SET USING RESNET-50 AS THE BACKBONE NETWORK.

Method	1/16	1/8	1/4	1/2
CCT [13]	64.2	70.7	73.2	75.3
GCT [31]	63.9	70.0	73.0	74.8
ST++ [16]	-	70.8	72.1	-
iMAS [9]	71.1	74.2	75.2	77.2
UniMatch [33]	71.5	73.9	74.6	76.8
CPS [3]	71.2	74.0	74.7	77.0
PSMT [29]	-	73.5	73.9	76.4
RankMatch [34]	71.7	74.7	75.0	77.3
ours	**72.8**	**75.6**	**75.7**	**77.5**

TABLE IV
ABLATION STUDY RESULTS OF THE DA AND FP MODULES ON THE 1/8 PARTITION OF THE PASCAL VOC 2012 DATASET.

DA	FP	mIOU (%)
		58.9 (supervised)
✓		72.7
	✓	72.4
✓	✓	73.8

TABLE V
IMPACT OF DIFFERENT NUMBERS OF STRONGLY AUGMENTED VIEWS ON MODEL PERFORMANCE UNDER DIFFERENT PARTITIONING SETTINGS.

Method	1/16 (92)	1/8 (183)	1/4 (366)	1/2 (732)
Single	67.4	72.2	75.5	77.4
Dual	68.0	72.7	76.0	77.6
Triple	67.8	72.3	75.8	77.1
ours	**68.5**	**73.8**	**76.6**	**78.2**

C. Ablations Studies

We conduct ablation studies on the 1/8 partition of PASCAL VOC 2012 and Cityscapes based on ResNet-101.

The effectiveness of our data augmentation. To validate the effectiveness of our method's components, we conducted ablation experiments on the PASCAL VOC 2012 dataset to evaluate the contributions of our dual adaptive augmentation at the image level (DA) and feature-level perturbation(FP) to model performance. As shown in table IV, when the DA component is introduced, mIoU increases to 72.7%. Additionally, when only the FP component is added (without DA), mIoU reaches 72.4%, indicating that feature-level perturbation enhances the robustness of the model. When both DA and FP components are included, mIoU further improves to 73.8%, achieving a 14.9% improvement over the supervised baseline. These results demonstrate the complementary nature of image-level and feature-level enhancement strategies. The experiments confirm that combining both image-level and feature-level augmentation strategies is essential for improving semi-supervised segmentation performance.

The diversity of perturbation strategies is essential. From the results, it is evident that the effectiveness of diverse perturbations is crucial. In our approach, we employ two types of image-level augmented views and one feature-level perturbation stream. To verify that increasing perturbation diversity is more effective than simply adding more augmented images, we designed a counterpart method using three types of image-level augmentations only. As shown in table V, our method consistently outperforms this counterpart under all partition protocols. This suggests that the performance gain stems not just from the number of views but from the diversity across different types of perturbation strategies.

The insertion position of feature perturbation is crucial. We compare the effects of perturbation applied at the encoder-decoder boundary and at the final classifier. As shown in table VI, applying perturbation at the encoder-decoder boundary (En-Decoder) performs better than applying it at the final classifier (Decoder).

Setting the confidence threshold. According to fig. 2, we

TABLE VI
ABLATION STUDY ON THE INSERTION LOCATIONS OF FEATURE PERTURBATION.

Location	92	183	366	732	1464
Decoder	65.0	70.9	75.6	76.4	79.3
En-Decoder (ours)	68.5	73.8	76.6	78.2	79.6

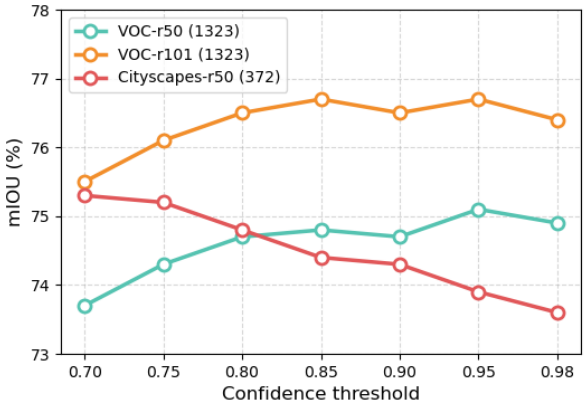

Fig. 2. Ablation study on different confidence threshold τ in our method on Pascal and Cityscapes datasets.

set $\tau = 0.95$ as the default setting for the VOC dataset, while $\tau = 0.7$ is used as the default setting for the CityScapes dataset. Due to the higher challenge of the CityScapes dataset, it requires a stronger discriminative ability. A higher confidence threshold may restrict the model from effectively learning from unlabeled samples. Therefore, a lower threshold is adopted for CityScapes to better facilitate semi-supervised training performance.

The necessity of independently separating image-level and feature-level perturbations. fig. 3 illustrates the impact of different perturbation strategies on model performance, verifying the necessity of independently separating image-level and feature-level perturbations. The Single strategy adopts a single mixed stream, integrating feature perturbations into image-level strong augmentations. The Double strategy employs a dual-mixed stream, integrating feature perturbations into two image-level augmentations. In contrast, the Ours method completely separates the two image-level augmentations from feature-level perturbations, forming three independent streams. The results indicate that the Ours method achieves the best performance across all labeled data partitions, demonstrating that independently modeling perturbations at different levels can enhance model stability and generalization.

Qualitative Results. As illustrated in fig. 4, our model produces more accurate segmentation of challenging objects such as motorcycles and humans compared to the baseline and iMAS, highlighting its superiority.

V. CONCLUSION

This study addresses two common challenges in semi-supervised semantic segmentation: the varying difficulty levels among unlabeled samples and the limited perturbation space in training. To overcome these limitations, we propose a Dual Adaptive Augmentation method and experimentally validate its effectiveness. First, we assess sample difficulty via a class-weighted Intersection over Union (IoU) metric, enabling dynamic adjustment of perturbation intensity according to each instance's learning demand. Second, we implement a dual-level perturbation mechanism that combines two strong image-level augmentations with stochastic feature-level perturbations, thereby expanding the diversity of the perturbation space. Experimental results demonstrate that these strategies operate synergistically, yielding significant improvements in segmentation performance and offering a more robust and adaptive solution for semi-supervised semantic segmentation.

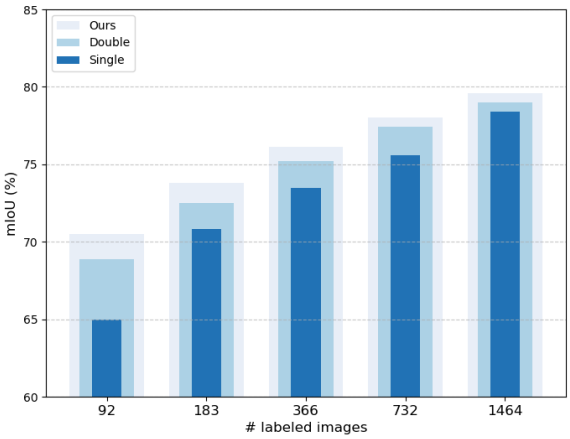

Fig. 3. Ablation Study on Different Perturbation Strategies.

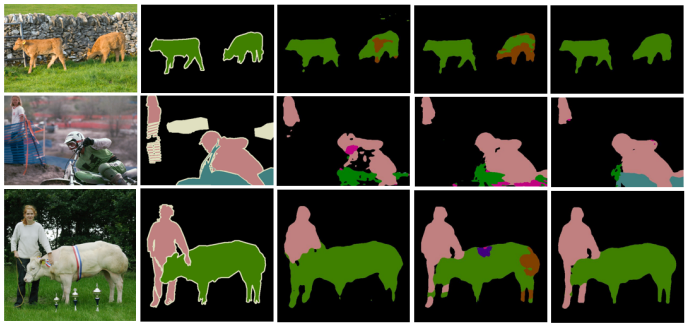

Fig. 4. This is our qualitative experimental result on the Pascal VOC 2012 dataset using 183 labeled images. (a) Original image, (b) Ground-Truth, (c) Segmentation results of the state-of-the-art method iMAS, (d) Segmentation by our proposed method.

REFERENCES

[1] Wei-Chih Hung, Yi-Hsuan Tsai, Yan-Ting Liou, Yu-Ting Lin, and Ming-Hsuan Yang, "Adversarial learning for semi-supervised semantic segmentation," in *Proceedings of the British Machine Vision Conference (BMVC)*, 2018.

[2] Shivansh Mittal, Maxim Tatarchenko, and Thomas Brox, "Semi-supervised semantic segmentation with high- and low-level consistency," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2019.

[3] Xiaokang Chen, Yuhui Yuan, Gang Zeng, and Jingdong Wang, "Semi-supervised semantic segmentation with cross pseudo supervision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[4] Geoffrey French, Timo Aila, Samuli Laine, and Michal Mackiewicz, "Semi-supervised semantic segmentation needs strong, high-dimensional perturbations," in *Proceedings of the British Machine Vision Conference (BMVC)*, 2020.

[5] Huaijin Hu, Fanwei Wei, Hua Hu, and Qiong Ye, "Semi-supervised semantic segmentation via adaptive equalization learning," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[6] Jianjian Yin, Yi Chen, Zhichao Zheng, Junsheng Zhou, and Yanhui Gu, "Uncertainty-participation context consistency learning for semi-supervised semantic segmentation," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, To appear.

[7] Hongyi Zhang, Moustapha Cisse, and Yann N. Dauphin, "mixup: Beyond empirical risk minimization," *arXiv preprint*, vol. arXiv:1710.09412, 2017.

[8] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, and Nicholas Carlini, "Fixmatch: Simplifying semi-supervised learning with consistency and confidence," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[9] Ziyu Zhao, Shuangjun Long, and Jianzhuang Pi, "Instance-specific and model-adaptive supervision for semi-supervised semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[10] Guangyuan Gui, Ziyu Zhao, Lu Qi, Li Zhou, Liang Wang, and Yinghuan Shi, "Improving barely supervised learning by discriminating unlabeled samples with super-class," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[11] Zhuangwei Hu, Shu Yang, Yunchao Wei, et al., "Semimatch: Semi-supervised learning with one labeled image per category," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[12] Xiaoyang Wang, Bingfeng Zhang, Limin Yu, and Jimin Xiao, "Hunting sparsity: Density-guided contrastive learning for semi-supervised semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[13] Yassine Ouali, Christophe Hudelot, and Mohamed Tami, "Semi-supervised semantic segmentation with cross-consistency training," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[14] Yingda Wang, Huan Wang, Yiming Shen, and Jingyuan Fei, "Semi-supervised semantic segmentation using unreliable pseudo-labels," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[15] Yuliang Zou, Zilei Zhang, Han Zhang, and Chong-Lin Li, "Pseudoseg: Designing pseudo labels for semantic segmentation," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

[16] Limin Yang, Wenjing Zhuo, Lu Qi, and Yinghuan Shi, "St++: Make self-training work better for semi-supervised semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[17] Jiawei Sung, Seong Joon Oh, and Minsu Cho, "Contextrast: Contextual contrastive learning for semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2024, pp. 12345–12354, IEEE.

[18] Yunhang Zhong, Bowen Yuan, Huaxin Wu, Zehao Yuan, Jinjun Peng, and Yu-Xiong Wang, "Pixel contrastive-consistent semi-supervised semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[19] Yuying Zhou, Han Xu, Wenxuan Zhang, Bingjie Gao, and Pheng-Ann Heng, "C3-semiseg: Contrastive semi-supervised segmentation via cross-set learning and dynamic class-balancing," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[20] Ekin D. Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V. Le, "Autoaugment: Learning augmentation policies from data," *arXiv preprint*, vol. arXiv:1805.09501, 2018.

[21] Ekin D. Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V. Le, "Randaugment: Practical automated data augmentation with a reduced search space," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2020.

[22] Mark Everingham, Samira M. A. Eslami, Luc Van Gool, and Christopher K. I. Williams, "The pascal visual object classes challenge: A retrospective," *International Journal of Computer Vision (IJCV)*, 2010.

[23] Marius Cordts, Mohamed Omran, Sebastian Ramos, and Timo Rehfeld, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[25] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, and Kai Li, "Imagenet: A large-scale hierarchical image database," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[26] Liang-Chieh Chen, Yi Zhu, George Papandreou, and Florian Schroff, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[27] Bharath Hariharan, Pablo Arbelaez, and Lubomir Bourdev, "Semantic contours from inverse detectors," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2011, pp. 991–998.

[28] Ruibin He, Jiawei Yang, and Xiaojuan Qi, "Re-distributing biased pseudo labels for semi-supervised semantic segmentation: A baseline investigation," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2021.

[29] Wei-Chih Hung, Yi-Hsuan Tsai, Yan-Ting Liou, and Yu-Ting Lin, "Adversarial learning for semi-supervised semantic segmentation," in *Proceedings of the British Machine Vision Conference (BMVC)*, 2018.

[30] Moshe Kimhi, "Semi-supervised semantic segmentation via marginal contextual information," *Transactions on Machine Learning Research*, jun 2024.

[31] Jingjing Ke, Deng-Ping Chen, Jun Qi, and Alan L. Yuille, "Guided collaborative training for pixel-wise semi-supervised learning," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[32] Xiaoyang Wang, Huihui Bai, and Yu, "Towards the uncharted: Density-descending feature perturbation for semi-supervised semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[33] Limin Yang, Wenjing Zhuo, Lu Qi, Yinghuan Shi, and Yang Gao, "Revisiting weak-to-strong consistency in semi-supervised semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[34] Huayu Mai, Rui Sun, Tianzhu Zhang, and Feng Wu, "Rankmatch: Exploring the better consistency regularization for semi-supervised semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 3391–3401.