Journal of Network Intelligence
Taiwan Ubiquitous Information
 ISSN 2414-8105 (Online)
Volume **10**, Number **1**, February **2025**

Multi-scale Inconsistent Semi-supervised Semantic Segmentation Architecture

Ye-Shen Guo, An-Hong Wang, Yuan-Zhi Gao, Zhi-Hong Li*, De-Biao Zhang

School of Electronic Information and Engineering
Taiyuan University of Science and Technology, Taiyuan 030024, China
guoyeshen@163.com, wah_ty@163.com, 542636407@qq.com
1994013@tyust.edu.cn, zhangdebiao@tyust.edu.cn

*Corresponding author: Zhi-Hong Li

Received February 15, 2024, revised June 30, 2024, accepted August 13, 2024.

Abstract. *Machine vision advancements suggest that multi-scale feature representations could enhance the performance of visual tasks. In contrast, existing neural networks maintain a singular scale approach based on a consistent teacher-student paradigm. Considering the deep semantic connections inherent to each pixel revealed by multi-scale attributes, and with the objective of reducing the parameterization scope of the entire neural network, this paper presents the design of a semi-supervised semantic segmentation network. This network cleverly integrates multi-scale modules and inconsistent policies tailored to the teacher-student architecture. By generating feature representations at different scales for the same feature map during the downsample process, the receptive field can be enlarged, resulting in a more detailed semantic understanding. Moreover, by adopting an approach in which the teacher and student models use inconsistency strategy, it is possible to mitigate the influence of random noise present in pseudo-labeling, which could otherwise lead to the student model's overfitting. Empirical results support that incorporating a multi-scale approach results in a performance increase of 1.65%, surpassing the established baseline. An additional improvement of 0.12% is seen upon integrating the inconsistency strategy, further confirming the efficacy of the proposed methods.*
Keywords: semi-supervised learning, semantic segmentation, multi-scale

1. **Introduction.** In the realm of machine vision, the purpose of the semantic segmentation task [1–4] is to imbue each pixel of an image with semantic labels. Although algorithms based on supervised learning for semantic segmentation are known to give notable results, the painstaking process of assigning detailed annotations to each pixel in large datasets is often prohibitively costly and remarkably time-consuming. Hence, the semi-supervised learning approach presents itself as an efficacious alternative to the dilemma of possessing a limited pool of meticulously labeled data alongside a copious amount of unlabeled data during the training phase. Semi-supervised learning endeavors to utilize the limited set of thoroughly labeled examples while simultaneously capitalizing on the abundant collection of unlabeled samples to train a model that approximates the performance of its supervised counterpart [5–7]. Therefore, in this paper, we use semi-supervised learning on the semantic segmentation task.

In the burgeoning domain of semi-supervised semantic segmentation, noteworthy advancements have been guided by the paradigms of entropy minimization, consistency regularization, and their synergistic integration. Among these, Xie et al. [6] employed a self-training approach within the framework of entropy minimization, initially training the

teacher model using the supervised approach. Subsequently, the trained teacher model was utilized to predict labels for the unlabeled data, and the generated pseudo-labels were incorporated into the training set for the re-training of the student model. Their work capitalized on a substantial volume of unlabeled data and, through iterative training of the teacher-student model, progressively included the unlabeled data within the training scope—thereby enhancing data utilization. Nonetheless, this approach does not circumvent the inherent complexities of computational demands and the meticulous tuning of hyperparameters, which require astute direction in practical applications. The architectural design of the Mean Teacher, as propounded by Tarvainen and Valpola [7], adheres to the principle of consistency constraints, postulating that the teacher model exhibits a greater stability relative to the student model. It therefore imposes consistency constraints from the student network to the teacher network on all unlabeled samples. Additionally, utilizing the mean teacher model for prediction smoothing helps reduce the risk of overfitting and improves the generalization ability of the model. However, this methodology necessitates a substantial increase in computational expenditure and temporal investment for model training, which can be taxing on available computational resources. Yang et al. [8]. revisited the potential of the classical self-training paradigm in the realm of semi-supervised semantic segmentation. Through the synthesis of the dual techniques of entropy reduction and consistency enforcement, they have advanced a strategy that, by merely introducing a selection of perturbations, markedly enhances performance. Considering the superlative outcomes this stratagem yields within the self-teaching algorithms, we have elected this succinct and potent methodology as our baseline.

1.1. **Realted work.**

1.1.1. *Semi-supervised learning.* Semi-supervised learning constitutes a paradigm in which a classifier is trained utilizing a voluminous collection of unannotated samples augmented by a scant aggregation of annotated samples, thereby surmounting the obstacle of an insufficiency in annotated data. In the vanguard of semi-supervised learning, a pair of methodologies has emerged to tackle this impediment: entropy minimization [6, 8] and consistency regularization [5, 7]. The antecedent encompasses the self-training algorithm, which generates pseudo-labels for the unlabelled datasets, subsequently amalgamated with manually annotated data to augment the training of the model. In contrast, the latter endeavors to engender stable and congruous prognostications for identical unlabeled data when subjected to a variety of perturbations. FixMatch [9], informed by the methodology of MixMatch, assimilates the merits of both stratagems whilst advancing further refinements. Subsequent endeavors, such as FlexMatch [10] and FreeMatch [11], interrogate the expedience of class-specific thresholding as a means to excise labels of diminished confidence.

1.1.2. *Semantic Segmentation.* Image semantic segmentation entails the execution of dense, pixel-level prognostications upon input visuals. With the rapid development of convolutional neural networks, significant progress has been made in various tasks within the field of computer vision, including image semantic segmentation, image classification [12], motion recognition [13], and video processing [14]. Long et al. [1]. delineated the Fully Convolutional Network (FCN), promulgating an architectural paradigm predicated on wholly convolutional networks that now pervades the semantic segmentation domain. Complementarily, the implementation of the Encoder-Decoder [15–17] architecture has proven propitious for mining more profound feature strata within networks. Chen et al. [3]. unveiled the Atrous Spatial Pyramid Pooling construct, a mechanism adept at ensnaring contextual nuances. Concurrently, Xu et al. [18]. have investigated the synergistic

deployment of Convolutional Neural Networks (CNNs) alongside Proportional-Integral-Derivative (PID) controllers, thus charting novel pathways in computational perception.

1.1.3. *Semi-supervised Semantic Segmentation.* In the initial work, semi-supervised semantic segmentation utilized Generative Adversarial Networks (GANs) [19] to differentiate between pseudo-labels and manual labels. Later, feasible ways to improve performance were explored from the domains of entropy minimization and consistency regularization. In recent experiments, the Cutmix [20] method has been shown to be effective for consistency regularization. Concurrently, U2PL [21] methodology, echoing the precepts of co-training, meticulously identifies pixels falling beneath a certain probabilistic demarcation as negative exemplars, thereby creating a dichotomy with their positive counterpart, CPS [22] proposes the use of two model branches for mutual supervision. PseudoSeg [23] makes further work to refine the mask based on FixMatch [9].

In the field of self-training methodologies, ST++ [8] employs the straightforward tactic of mitigating the tendency of student models to overfit on noisy pseudo-labels by integrating suitable data augmentation efforts, substantially elevating network efficacy. Nevertheless, it is acknowledged that this endeavor does not confront the quintessential challenges inherent in the task of semantic segmentation. We have resolved to assimilate the ASP module, conceived to expand the receptive field, into the underlying architecture, thus promoting a synthesis of context and augmenting the semantic granularity of convolutional neural networks. Moreover, this study explores innovative methodologies to mitigate the impact of noise within pseudo-labels on the student network performance. Additionally, it seeks to refine the network by reducing its parameter complexity, thereby enhancing computational efficiency.

1.2. **Motivation and contribution.** In harnessing the profound capabilities of convolutional neural networks for semantic segmentation and dense predictive tasks, we observe that as the network architecture subjects the image to sampling reductions, there ensues a corresponding decline in resolution, resulting in the loss of exquisitely fine details. Given that current benchmark methodologies inadequately address this critical issue, this paper adopts a multiscale approach and proposes an ASP module, predicated on Atrous convolution, to encapsulate rich details. To counteract the potential overfitting to noise inherent in the self-training algorithm—stemming from the student model's adoption of pseudo-labels engendered by the teacher model—an intentional discrepancy between the teacher and student models has been orchestrated. Such a strategy diminishes not only the dependence on reference data but also augments the performance, thereby exceeding the capabilities of the original framework. The contributions of this paper can be itemized as follows:

(1) The devised architecture enhances the precision of feature delineation and modulates the granularity of feature resolution via the implementation of Atrous Convolutions across divergent branches. This approach provides a macroscopic augmentation of the network's receptive field concerning the input imagery and coalesces global data whilst interlinking contextual elements.

(2) Furthermore, we have devised an innovative Teacher-Student Model Inconsistent Strategy aimed at attenuating the perturbations introduced by noise in pseudo-labeling processes. This strategy is curbing computational expenditure and temporal investments while catalyzing further enhancement in performance metrics.

2. **Overview of the modeling framework.**

2.1. **Network architecture.** The semi-supervised semantic segmentation network architecture for inconsistent multi-scale modular teacher-student network proposed in this paper follows the traditional self-training approach for training. As shown in Figure 1, firstly, the images x_i and labels y_i in the labeled dataset $D^l = \{(x_i, y_i)\}_{i=1}^{M}$ are input into the model to train the teacher model **T** using cross-entropy loss L_s.Within this procedure, we duly consider the implications of augmenting the receptive field upon the feature extraction capabilities at various scales. Consequently, the ASP module is plugged into the Encoder part of the model as a serial connection. Second, the teacher model **T** is used to generate pseudo-labels v_i by predicting the unlabeled images u_i in the unlabeled dataset $D^u = \{u_i\}_{i=1}^{N}$, and then the unlabeled data is integrated into the pseudo-labeled dataset $\widehat{D}^u = \{(u_i, v_i)\}_{i=1}^{N}$. Finally, during the training of the student model **S**, in order to minimize the interference of the noisy pseudo-labels on the student model **S** and avoid the coupling problem of making similar predictions for the same inputs, we adopt the original structure for training. The labeled dataset $D^l = \{(x_i, y_i)\}_{i=1}^{M}$ and unlabeled dataset $\widehat{D}^u = \{(u_i, v_i)\}_{i=1}^{N}$ are combined, and the $\{(x_k, y_k)\}_{k=1}^{B} \subset \left(D^l \cup \widehat{D}^u\right)$ is fed into the model to train the student model **S** using loss L_u in Equation (1). In addition, we only make one iteration in the above training process, thereby substantially diminishing both the requisite computational resources and the temporal investment for training. The loss objective is formalized as follows:

$$L_u = H\left(y_k, \boldsymbol{S}\left(x_k\right)\right) \tag{1}$$

where H denotes entropy minimization between student and label.

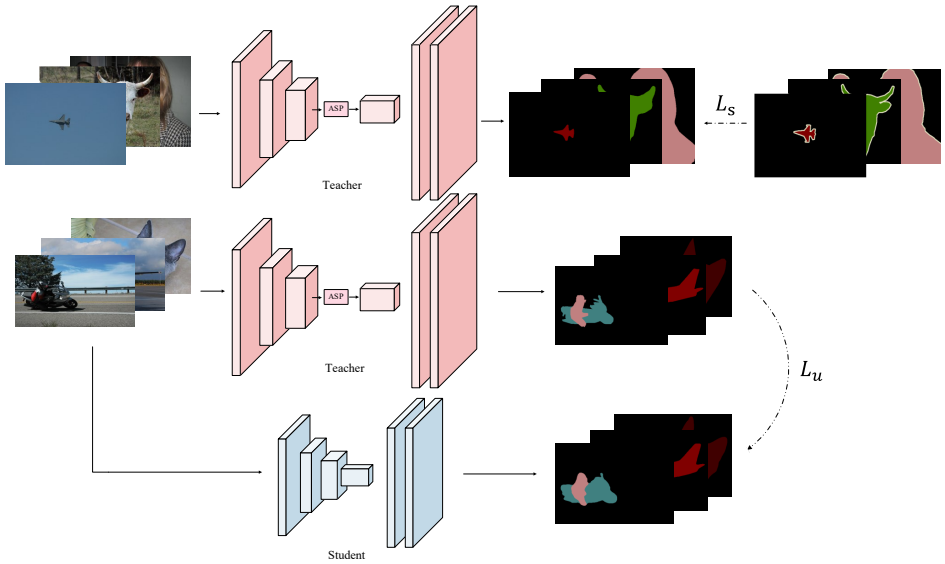

FIGURE 1. Overall network structure figure

2.2. **ASP module.** In pursuit of eliciting deeper semantic interrelations within the input features, we draw inspiration from the scholarly contributions elucidated in [2] and introduce the ASP module, which possesses the aptitude to augment the network model's receptive expanse for the feature maps, thereby amplifying the model's proficiency in contextual correlation and the synthesis of dimensional data. As shown in Figure 2, the ASP module employs an array of concurrent branches to apprehend feature representations amidst divergent receptive domains for an identical input feature map, thereby

augmenting the semantic interconnectivity among individual pixels. Specifically, as shown in Equation (2), for a deep feature expression x in the downsample layer, we input it into five branches in parallel. In the first branch, x is fed into the Conv1x1 convolution to execute a dimensionality reduction while preserving its spatial magnitude, subsequently undergoing batch normalization and ReLU layer processes to calibrate the feature map dimensions to the predetermined metrics. The ensuing branches, numbered second through fourth, employ the parameter $\tau = [12, 24, 36]$ to concurrently enact Atrous Convolution operations on the input x. The distinct branches yield feature mappings across a variety of scales attributable to their respective receptive fields. This engenders a more multifaceted portrayal of the input feature x, thereby permitting the neural network to assimilate more intricate feature information. Then, in the remaining branches, the global average pooling and the convolution of Conv1*1 are used to obtain global information about the input x through the BN and ReLU layers. Subsequently, this global feature is restored to a preordained dimensionality by implementing the bilinear interpolation of the feature maps. Finally, the different scale feature maps obtained from the branching are stacked, and the stacked features are convolved by Conv3*3 and then dimensionally adjusted before integrating all the features generated by the cascade structure by Conv1*1, the formula is expressed as follows:

$$\begin{cases} \boldsymbol{F}_{\text{conv } 1\times1} = \boldsymbol{f}_{\text{conv } 1\times1}(\boldsymbol{x}) \\ \boldsymbol{F}_{\boldsymbol{r}=12} = \boldsymbol{f}_{\boldsymbol{r}=12}(\boldsymbol{x}) \\ \boldsymbol{F}_{\boldsymbol{r}=24} = \boldsymbol{f}_{\boldsymbol{r}=24}(\boldsymbol{x}) \\ \boldsymbol{F}_{\boldsymbol{r}=36} = \boldsymbol{f}_{\boldsymbol{r}=36}(\boldsymbol{x}) \\ \boldsymbol{F}_{\text{conv } 3\times3} = \boldsymbol{f}_{\text{conv } 3\times3}(\boldsymbol{x}) \end{cases} \tag{2}$$
$$\boldsymbol{F}_{\text{out}} = \textbf{concat}\left(\boldsymbol{F}_{\text{conv } 1\times1}, \boldsymbol{F}_{\boldsymbol{r}=12}, \boldsymbol{F}_{\boldsymbol{r}=24}, \boldsymbol{F}_{\boldsymbol{r}=36}, \boldsymbol{F}_{\text{conv } 3\times3}\right)$$
$$\boldsymbol{F}_{\text{final}} = \boldsymbol{f}_{\text{conv } 1\times1}\left(\boldsymbol{f}_{\text{conv } 3\times3}\left(\boldsymbol{F}_{\text{out}}\right)\right)$$

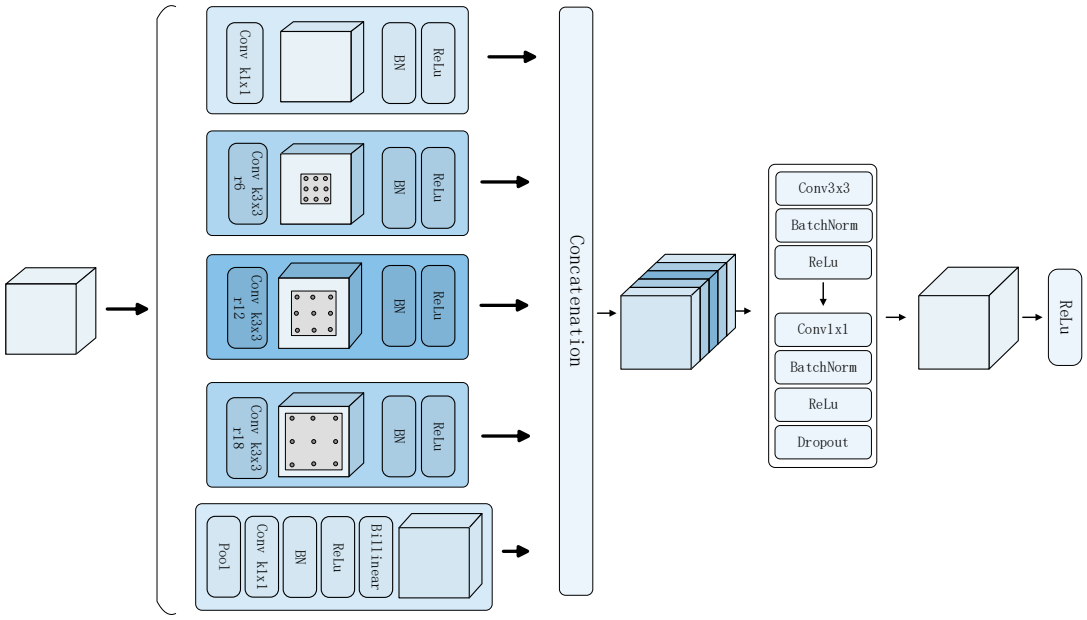

FIGURE 2. ASP module structure diagram

where $f_{r=i}$ denotes the Atrous Convolution operation with expansion factor r equal to i, $f_{conv1*1}(*)$ and $f_{conv3*3}(*)$ denote the convolution operation with convolution kernel

1*1 and convolution kernel 3*3, respectively, and $concat(*)$ denotes the splicing operation between features.

Given that this module amalgamates Atrous Convolution with Depthwise Separable Convolution, it markedly enhances computational velocity whilst ensuring performance integrity. In particular, for Atrous Convolution, consider a 2D data input feature map x, for each position i and a filter ω on the output feature map y.

$$y[i] = \sum x[i + r \cdot k]w[k] \tag{3}$$

where k is the size of the convolution kernel, and we refer the interested reader to [24] for more details. For computing the sensory field there is the following formula:

$$\boldsymbol{RF} = \frac{\boldsymbol{N} + \boldsymbol{P} \times 2 - \boldsymbol{K}}{\boldsymbol{S}} + 1 \tag{4}$$

where **N** is the size of the input image, **S** is the step size and **P** is the padding. We decide to take different parameters to get different scales of information by controlling the range of the sensory field as well as the size and dimension of the output features according to the formulation.

3. **Inconsistent Strategies(IS).** In traditional self-training based algorithms, the teacher model **T** is generally used to predict the unlabeled images u_i in the unlabeled dataset $D^u = \{u_i\}_{i=1}^N$, and the prediction results with prediction probability greater than a certain threshold ι are used as pseudo-labels v_i, and then the student model **S** is trained, and the process is usually repeated for a fixed number of iterations. However, in this type of operation, since the teacher model **T** cannot predict the unlabeled image very accurately, the generated pseudo-label v_i is likely to have noise in it, which causes the student model **S** to learn the noise incorrectly. Therefore, no threshold is set and only one iteration is done in our algorithm.

Furthermore, conventional self-training algorithms embrace an instructional paradigm in which homologous network architectures are employed for both the teacher model **T** and the student model **S**, the re-trained **S** is enforced to learn the pseudo labels from **T** in a supervised manner. However, there is a serious coupling problem between them, i.e., they make similar predictions on the same inputs [8]. In view of this situation, we have adopted a strategy that is exploratory in nature: We have chosen to conduct experiments with a neural network that is inconsistent between the teacher model **T** and the student model **S**. The aim is to alleviate the coupling problem described above and provide a solution to improve the performance of the model. Specifically, during the training of the teacher model using the supervised approach, we considered the positive effect of increasing the sensory field on the extraction of features by adding an ASP module to the Encoder part of the teacher model. Conversely, to mitigate the interference caused by noisy pseudo-labels on the student model **S** and to avert the coupling problem, we refrain from incorporating the ASP module, instead opting for training using the original structure. The experimental results demonstrate that our approach significantly improves the performance and decreases the number of parameters required, while simultaneously having a beneficial effect on the training process of the student model.

4. Experimentation and Analysis.

4.1. Experimental environment.

TABLE 1. Performance Comparison Table

Method	1/16	1/8	1/4
SupOnly	64.8	68.3	70.5
ECS [28]	-	70.2	72.6
DCC [29]	70.1	71.4	72.8
ST++	70.04	71.71	72.37
Ours (Only ASP)	71.69	73.23	73.88
Ours (ASP+IS)	**71.81**	**73.33**	**74.24**

4.1.1. *Datasets.* In the present study, we have employed the Pascal VOC 2012 dataset [25], which encompasses the tasks of Object Classification, Object Detection, Object Segmentation, and Action Classification. Within the ambit of the segmentation endeavor, a corpus of 2913 images have been allocated for the processes of training and validation, while a separate set of 1456 images have been designated for testing purposes. To augment the robustness of the training dataset, annotations encapsulating the degree of accuracy have been integrated into the SBD dataset [26].

4.1.2. *Network structure.* In prior research, Resnet [27] has revealed the concept of residual learning—an approach that adeptly tackles the ubiquitous challenges of gradient vanishing and explosion encountered in the training regimens of profound neural structures. Its sophisticated and robust architecture has markedly accelerated progress in the domain of deep neural computation. Consequently, we have adopted the architectural model of Resnet as the cornerstone upon which our model is predicated.

4.1.3. *Implementation details.* In the domain of network training, our experiments predominantly employ a quartet of NVIDIA GeForce GTX 1080 graphics cards. The batch magnitude for the training construct is 16, whilst the parameter - crop dimension is defined at a resolution of 321 by 321 pixels. We take SGD optimizer for the training task and use variable learning rate dynamically during the training process, other settings refer to the parameters of the baseline.

4.2. **Comparative Experiments.**

4.2.1. *Performance experiments.* In consonance with prior scholarly endeavors, our investigation also applies its methodologies to an enhanced iteration of the Pascal VOC 2012 dataset, which encompasses a total of 10,582 distinct images. Our evaluative criteria employ progressively larger fractions of the dataset, namely 1/16, 1/8, and 1/4, to ascertain the efficacy of our approach. Table 1 delineates the outcomes seized from the induction of ASP modules in series to the network. Procuring performance metrics transcending the established baseline across the data subsets of 1/4, 1/8, and 1/16 in magnitude, and it significantly surpasses that of alternative techniques. The integration of the Inconsistent Strategies (IS) fostered additional enhancements to the already augmented results.

4.2.2. *Visualization charts.* We have visualized the generated prediction map, as shown in Figure 3. We conducted experiments on 1/4 of the dataset, and compared with the baseline, our method reveals finer-grained semantic information in detail. Where, column (a) represents the input image column (b) represents the prediction map generated by the baseline column (c) represents the prediction map generated by our method column (d) represents the Ground Truth. The last row of visualized images also indicates that our

method is less affected by noise and is capable of accurately segmenting pixels belonging to the same category.

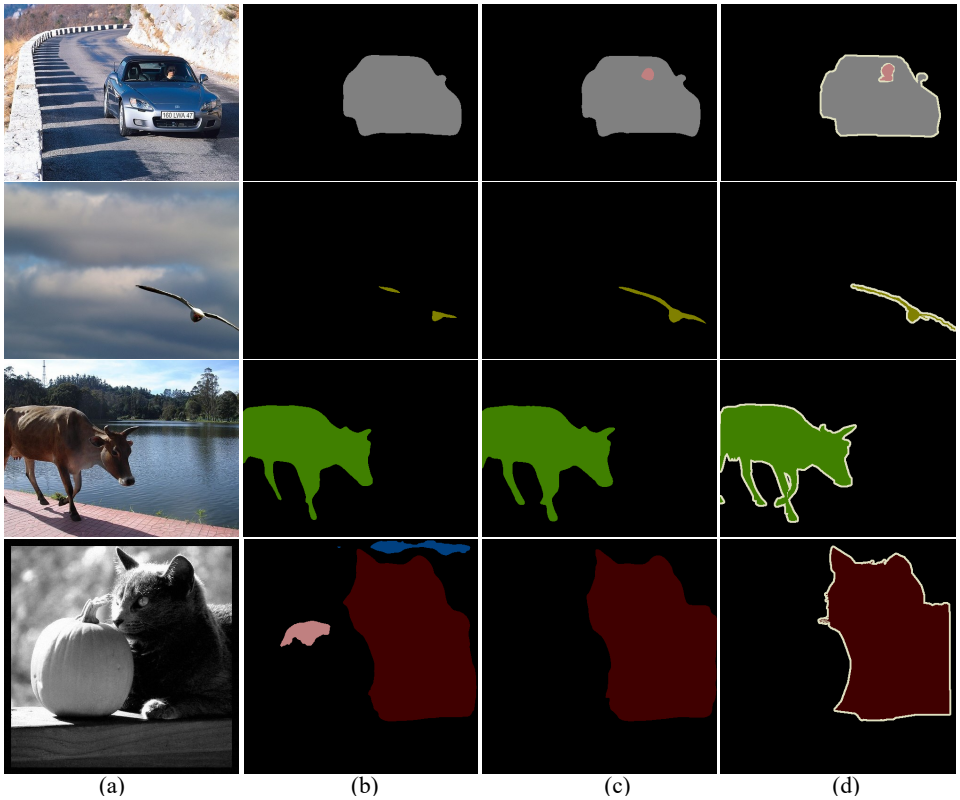

FIGURE 3. Visualization results chart

4.3. **Comparative Experiments.** In the pursuit of ascertaining the intricacy of computation, we employ the methodology wherein the parameter total, denoted as P_t, represents the aggregate of the quantity of parameters P_T within the teacher network and the corresponding quantity of parameters P_S within the student network, specifically.

$$P_t = P_T + P_S \tag{5}$$

As shown in Table 2, we conduct experiments on the 1/16 dataset separately, the number of parameters in the overall network rises significantly after the addition of the ASP module, to make a lightweight network, we use an inconsistent strategy to reduce the number of parameters while further improving the performance, and the results of this experiment demonstrate the effectiveness of our method in terms of the number of parameters.

TABLE 2. Explanatory table of parameters

Method	P_t(M)	mIoU(%)
ST++	80.8	70.04
Ours (ASP)	119	71.69
Ours (ASP+IS)	99.9	71.81

4.4. **Ablation experiments.** As shown in Table 3, we conducted ablation studies encompassing all dataset subcategories, thereby substantiating the efficacy of our work with respect to the employment or omission of the ASP module along with the inconsistency strategy.

TABLE 3. Table of ablation experiments

ASP	IS	1/16	1/8	1/4
		70.04	71.71	73.37
√		71.69	73.23	73.88
√	√	71.81	73.33	74.24

5. **Conclusion.** We propose a multiscale teacher-student network inconsistent network structure, aspiring for the network to assimilate copious information in the downsample segment via the multiscale module. Recognizing the potential adverse impact of noise induced by pseudo-labeling on student network training, we employ an inconsistent strategy. First, we train the teacher model incorporating the multiscale module and employ it to generate pseudo-labels for predictions on unlabeled images. Subsequently, we train the student model without the multiscale module using the combined dataset comprising pseudo-labels and labeled labels. Within this endeavor, we have reevaluated self-training algorithms and introduced the inconsistent strategy described above, with aspirations of advancing to surmount forthcoming challenges.

Acknowledgment. This work has been supported in part by National Natural Science Foundation of China (U23A20314, 62072325), Industrial Vision Application of Shanxi Provincial Technology Innovation Center (IVA-SXTIC2022), Shanxi Key Core Technology & Common Technology Research and Development Project (20201102011), Shanxi S&T Major Project (20191102010), Shanxi University S&T Achievements Transformation Cultivation Project (20191042), Shanxi S&T Achievements Transformation Project (201804D131035).

REFERENCES

[1] 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.

[2] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017.

[3] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 801–818.

[4] 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.

[5] 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.

[6] Q. Xie, M.-T. Luong, E. Hovy, and Q. V. Le, "Self-training with noisy student improves imagenet classification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 687–10 698.

[7] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," *Advances in Neural Information Processing Systems*, vol. 30, pp. 1–10, 2017.

[8] 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.

[9] K. Sohn, D. Berthelot, N. Carlini, Z. Zhang, H. Zhang, C. A. Raffel, E. D. Cubuk, A. Kurakin, and C.-L. Li, "Fixmatch: Simplifying semi-supervised learning with consistency and confidence," *Advances in Neural Information Processing Systems*, vol. 33, pp. 596–608, 2020.

[10] B. Zhang, Y. Wang, W. Hou, H. Wu, J. Wang, M. Okumura, and T. Shinozaki, "Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling," *Advances in Neural Information Processing Systems*, vol. 34, pp. 18 408–18 419, 2021.

[11] Y. Wang, H. Chen, Q. Heng, W. Hou, Y. Fan, Z. Wu, J. Wang, M. Savvides, T. Shinozaki, B. Raj *et al.*, "Freematch: Self-adaptive thresholding for semi-supervised learning," *arXiv preprint arXiv:2205.07246*, 2022.

[12] T.-Y. Wu, H. Li, S. Kumari, and C.-M. Chen, "A spectral convolutional neural network model based on adaptive fick's law for hyperspectral image classification," *Computers, Materials & Continua*, vol. 79, no. 1, 2024.

[13] F. Zhang, T.-Y. Wu, J.-S. Pan, G. Ding, and Z. Li, "Human motion recognition based on svm in vr art media interaction environment," *Human-centric Computing and Information Sciences*, vol. 9, no. 1, p. 40, 2019.

[14] F. Zhang, T.-Y. Wu, and G. Zheng, "Video salient region detection model based on wavelet transform and feature comparison," *EURASIP Journal on Image and Video Processing*, vol. 2019, no. 1, p. 58, 2019.

[15] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*. Springer, 2015, pp. 234–241.

[16] B. Cheng, L.-C. Chen, Y. Wei, Y. Zhu, Z. Huang, J. Xiong, T. S. Huang, W.-M. Hwu, and H. Shi, "Spgnet: Semantic prediction guidance for scene parsing," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 5218–5228.

[17] D. Mehta, A. Skliar, H. Ben Yahia, S. Borse, F. Porikli, A. Habibian, and T. Blankevoort, "Simple and efficient architectures for semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 2628–2636.

[18] J. Xu, Z. Xiong, and S. P. Bhattacharyya, "Pidnet: A real-time semantic segmentation network inspired by pid controllers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 19 529–19 539.

[19] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in Neural Information Processing Systems*, vol. 27, pp. 1–9, 2014.

[20] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, "Cutmix: Regularization strategy to train strong classifiers with localizable features," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 6023–6032.

[21] Y. Wang, H. Wang, Y. Shen, J. Fei, W. Li, G. Jin, L. Wu, R. Zhao, and X. Le, "Semi-supervised semantic segmentation using unreliable pseudo-labels," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4248–4257.

[22] X. Chen, Y. Yuan, G. Zeng, and J. Wang, "Semi-supervised semantic segmentation with cross pseudo supervision," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 2613–2622.

[23] Y. Zou, Z. Zhang, H. Zhang, C.-L. Li, X. Bian, J.-B. Huang, and T. Pfister, "Pseudoseg: Designing pseudo labels for semantic segmentation," *arXiv preprint arXiv:2010.09713*, 2020.

[24] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[25] M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge: A retrospective," *International Journal of Computer Vision*, vol. 111, pp. 98–136, 2015.

[26] B. Hariharan, P. Arbeláez, L. Bourdev, S. Maji, and J. Malik, "Semantic contours from inverse detectors," in *2011 International Conference on Computer Vision*. IEEE, 2011, pp. 991–998.

[27] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[28] 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.

[29] X. Lai, Z. Tian, L. Jiang, S. Liu, H. Zhao, L. Wang, and J. Jia, "Semi-supervised semantic segmentation with directional context-aware consistency," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1205–1214.