Pseudo Labeling Methods for Semi-Supervised Semantic Segmentation: A Review and Future Perspectives

Lingyan Ran, Yali Li, Guoqiang Liang, Yanning Zhang, *Senior Member, IEEE*

***Abstract*—Semantic segmentation is a fundamental task in computer vision and finds extensive applications in scene understanding, medical image analysis, and remote sensing. With the advent of deep learning, significant advancements have been made in segmentation tasks. However, deep learning models require a substantial amount of labeled data for training, and accurately annotating datasets is labor-intensive and costly. Recently, numerous studies have explored the semantic segmentation task through the lens of semi-supervised learning, with the pseudo-labeling (PL) method emerging as a straightforward and widely applicable approach. This paper provides a comprehensive review and analysis of various PL methods and their applications in semi-supervised semantic segmentation (SSSS) from multiple angles. Initially, it captures the essence of individual model self-training and the collaborative training of multiple models from a model-centric viewpoint. Next, it explores strategies for refining or dismissing unreliable methods. Then, it categorizes techniques for addressing noisy PL data and inspects improvements in PL methods from the perspective of data augmentation. It further provides insights into optimization strategies. Furthermore, it examines PL methods from an application-oriented standpoint, such as in medical image segmentation and remote sensing image segmentation. Lastly, this paper evaluates the performance of cutting-edge methods on public datasets and concludes by discussing the challenges and potential directions for future research.**

***Index Terms*—Semi-supervised semantic segmentation, pseudo-labeling, semi-supervised learning**

I. Introduction

SEMANTIC segmentation is a traditional and significant area within computer vision, focusing on classifying all pixels in an image based on their semantic content. In recent years, many successful studies have emerged in this field, with diverse applications in specific domains, such as natural image [1]–[3], medical image [4]–[6], and remote sensing image [7]–[10] analysis, driving scene segmentation [11]–[13], and point cloud segmentation [14]–[16].

The integration of deep learning has greatly improved the efficiency of semantic segmentation tasks. However, deep learning requires a large amount of labeled data to train effective models. Despite the ease of acquiring vast amounts of raw data nowadays, manually labeling images pixel by pixel is arduous and time-intensive and requires considerable manual effort. Utilizing the Cityscape dataset [17] as a representative case, annotation, and quality control required more than 1.5h on average for a single image. Consequently, effectively training models with insufficient annotations for open-world applications poses a significant challenge. To address this issue, researchers have explored integrating methods such as weakly supervised learning [18]–[22] or unsupervised learning [23]–[27]. Generally, semi-supervised learning (SSL) [28]–[30] provides an advantage by leveraging both labeled and unlabeled datasets.

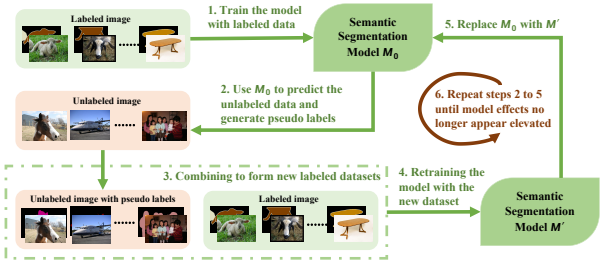


Fig. 1: The plainest process of SSSS with the pseudo-labeling method. The acquisition and qualification of pseudo-labels are the main focus of the framework.

Semi-supervised semantic segmentation (SSSS) aims to develop a model that can effectively utilize limited labeled data while extracting valuable insights from a vast amount of unlabeled data. With the rapid advancement of deep learning models in visual and linguistic fields, numerous groundbreaking solutions have been introduced to tackle SSSS tasks. In Fig. 2, a thorough overview of the development trends and future directions for SSSS is elaborated. These solutions can be broadly categorized into pseudo labeling (PL) [31]–[33], consistency regularization [34]–[36], adversarial training [37], [38], contrastive learning [39]–[42], and hybrid approaches [43]–[45]. Among these methods, pseudo-labeling is a well-established technique, which is increasingly promising due to its stability, interpretability, and ease of implementation. It was first introduced in [46] and has gained traction in recent computer vision research, including image classification [47], object detection [48], and semantic segmentation [49]. In addition, the PL has shown potential in improving model

Manuscript received 6 June 2024; revised 19 October 2024; accepted 25 November 2024. Date of publication XX November 2024; date of current version 25 November 2024.

This work is supported in part by the National Natural Science Foundation of China(62476226), Natural Science Basic Research Program of Shaanxi (2024JC-YBQN-0719), Natural Science Foundation of NingBo (2023J262). *(Corresponding author: Guoqiang Liang.)*

Lingyan Ran, Yali Li, Guoqiang Liang, and Yanning Zhang are with the Shaanxi Provincial Key Laboratory of Speech and Image Information Processing, and the National Engineering Laboratory for Integrated Aerospace-GroundOcean Big Data Application Technology, School of Computer Science, Northwestern Polytechnical University, Xi'an 710072, China.

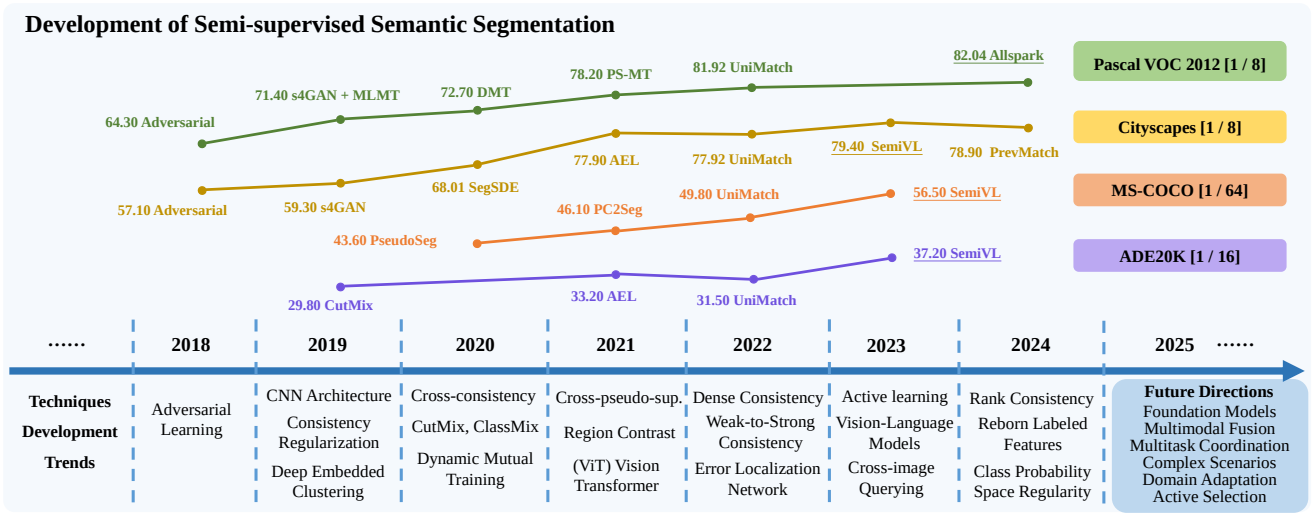


Fig. 2: The development trends of semi-supervised semantic segmentation. In the upper half, we provide solutions for SOTA retrieval performance on four more commonly used benchmark datasets from 2018 to 2024, respectively (the proportion in the box is the proportion of labeled data). As for the second half, we provide an overview of relevant technology trends in recent years and point out possible future research directions (best viewed in colors).

performance in other domains, including medical imaging [50], remote sensing [51], natural language processing [52], autonomous driving [53], speech recognition [54], and action recognition [55], where large-scale labeled data may be limited. By utilizing unlabeled data, PL techniques can develop more robust and versatile models across different tasks. The detailed process of PL for SSSS is illustrated in Fig. 1. Several reviews have been conducted on subjects like SSL [56] and SSSS [57], [58], yet comprehensive summaries and analyses for PL methods are still missing. The PL offers significant advantages in leveraging large amounts of unlabelled data, but there are persistent challenges, such as ensuring the quality and accuracy of generated pseudo-labels and addressing the noise and overfitting issues [46], [59] that pseudo-labels might introduce. Given this context, we are driven to perform a current and thorough survey to encapsulate the latest solutions in PL for SSSS, examine existing challenges, spark new research ideas, explore integration with other techniques, and discuss future directions to encourage this technology's broad adoption further.

We present a comprehensive overview of this survey in Fig. 3. The typical PL solutions for SSSS are categorized into four sections: model structure, pseudo-label refinement, data enhancement, and optimization improvement. Specifically, the model structure can be broadly classified into single-model and multi-model frameworks. Pseudo-label refinement is carried out to mitigate confirmation bias [60] and improve the quality of pseudo-labels, which is differentiated according to whether the pseudo-labels are updated or not. The optimization improvement section primarily encompasses loss function, data enhancement, label-data utilization, and pseudo-labels utilization. We detail and elaborate on the main contributions of various approaches from these three perspectives, with the specific components of each section depicted in Fig. 4.

Following this, we demonstrate the use of PL in various SSSS domains, such as medical imaging and satellite remote sensing. Next, we offer a detailed comparison and analysis of the current SOTA solutions through quantitative and qualitative experiments, aiming to provide valuable insights for researchers in related fields. In summary, we aim for this review to assist researchers in swiftly understanding the recent advancements and prospects of PL in SSSS.

Our main contributions are summarized as follows:

- A comprehensive review of the solutions of PL methods in SSSS is presented, with a focus on organizing the relationships and differences between them and discussing the advantages and limitations of each technique.
- We provide an extensive evaluation of existing PL methods by qualitative and quantitative experimental means.
- The emerging technologies available in SSSS are analyzed and summarized, and current challenges and potential advances in PL methods for SSSS are elaborated and discussed.

II. Pseudo-label Methods for SSSS

Numerous applications [50], [51], [61]–[64] have widely embraced pseudo-labels since their inception, offering the significant benefit of leveraging unlabeled data and enhancing model performance [65]–[67]. Nevertheless, they are constrained by noise and errors [60]. The decision of how to utilize pseudo-labels to mitigate model bias is also a major consideration. This section will provide an overview of recent studies that employ various strategies to enhance performance, improve stability, and broaden the scope of applications. In Table I, a summary of techniques examined in this review is classified according to their primary contributions.

A. Overall Problem Definition

Within the framework of SSSS, we possess a labeled dataset $D_l = \{(x_l, y_l)\}^p$, which contains p samples with associated

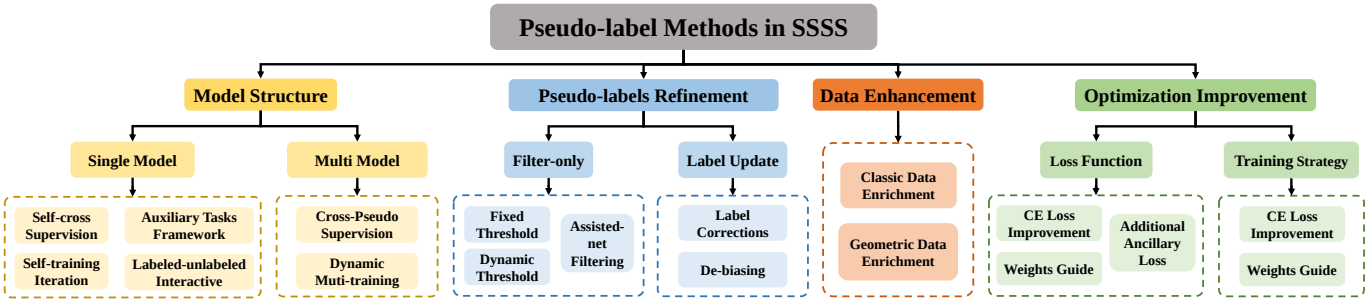


Fig. 3: Overall depiction of the logical framework in this survey (best viewed in colors).

labels, and an unlabeled dataset $D_u = \{x_u\}^q$, comprising q images without annotations, where $q >> p$. Our goal is to develop a deep model M that can minimize the training loss by taking into account both the labeled and unlabeled datasets.

As illustrated in Fig. 1, the vanilla PL approach involves the following steps: In the beginning, we train the initial model M_0 on the labeled dataset D_l utilizing the cross-entropy loss L^l. This training phase produces a pseudo-labeled dataset $\tilde{D_u} = \{(x_u, M_0(x_u))\}^q$ for the unlabeled dataset D_u, where $M_0(x_u)$ denotes the pseudo-labels of x_u. Next, we merge the labeled dataset $D_l = (x_i, y_i)$ with the pseudo-labeled dataset $\tilde{D_u}$ to create a comprehensive dataset $D = (D_l \cup \tilde{D_u})$. Finally, we train a new model M using the complete dataset D to replace M_0. The loss function $\mathcal{L}$ is composed of two components: $\mathcal{L}^{\mathrm{l}}$, which denotes the loss on the labeled dataset, and $\mathcal{L}^{\mathrm{u}}$, which signifies the loss on the unlabeled dataset. The total loss $\mathcal{L}$ can be expressed as:

$$\mathcal{L} = \lambda_l \mathcal{L}^{\mathrm{l}} + \lambda_u \mathcal{L}^{\mathrm{u}}, \tag{1}$$

where λ_l and λ_u represent a hyper-parameter that balances the trade-off, this can be set to a constant value in advance or dynamically modified throughout the training phase. The $\mathcal{L}^{\mathrm{l}}$ and $\mathcal{L}^{\mathrm{u}}$ can be expressed as:

$$\mathcal{L}^{\mathrm{l}} = -\frac{1}{\mathrm{p}} \sum_{l=1}^{\mathrm{p}} \sum_{c=1}^{C} y_{l,c} \log(p_{l,c}(x_l; \theta)) \tag{2}$$

$$\mathcal{L}^{\mathrm{u}} = -\frac{1}{\mathrm{q}} \sum_{u=1}^{\mathrm{q}} \sum_{c=1}^{C} \tilde{y}_{u,c} \log(p_{u,c}(x_u; \theta)) \tag{3}$$

where $p_{l,c}(x_l; \theta)$ and $p_{u,c}(x_u; \theta)$ is the model's predicted probability that the sample x_l and x_u belongs to category c, respectively. C is the number of categories. $y_{l,c}$ is the true corresponding to sample x_l with category label (one-hot coding). $\tilde{y}_{u,c}$ is the pseudo-label of x_u (which can be either hard or soft) generated by the model, usually selected by the maximum probability category output by the model. This straightforward procedure can be iteratively executed to continually enhance the quality of the generated pseudo-labels.

In the case of complex and challenging scenarios, simply generating solo hot labels and discarding low-confidence signals may result in a loss of information, leaving the training in a sub-optimal trap. Some of the accompanying work [60] introduced soft pseudo-labels, whereby a certain number of high-probability categories are retained as labels based on the pixel's prediction rather than a single category, all of which will be involved in the training, a more common practice [68], [69] is to combine soft and hard labels for training. Similarly, there is some work [70], [71] that soft-supervises the model directly using all the additional information retained in the soft output (soft confidence).

The process and performance of PL can be significantly influenced by the generation [72], [73], selection [74]–[76], and improvement [77]–[79] of the pseudo-labels due to the unexpected distribution gap [60] and the unsatisfactory performance [80] of pre-trained models.

B. Model Structure Perspective

The significance of architecture in deep learning cannot be overstated as it establishes the framework and layout of the neural networks employed in these models. The architecture plays a crucial role in determining how the network manipulates and analyzes input features, thereby impacting the model's capacity to learn and generate precise predictions. Selecting an appropriate architecture is a vital aspect of constructing a successful deep-learning model. In this subsection, we examine papers about single-model families and collaborative mutual-model families.

1) Single model based Methods:

Early Emergence: The concept of PL [46] was first proposed by Lee, who used a self-training technique [81], suggesting that the labels with a higher prediction probability from unlabeled data should be selected as pseudo-labels. The core idea of this approach is to generate pseudo-labels from the prediction results of the network, thus extending the training set with unlabeled data. However, the fluctuation and instability in the quality of pseudo-labels make the early methods face many challenges, such as the noise in the pseudo-labels may lead to degradation of the model performance [82], [83]. To improve SSL, some proposed Teacher-Student Framework [64], [84], [85]. In this case, the teacher model is updated by the exponential moving average (EMA) [84] of the student model, which consistently generates more consistent pseudo-labels. The student model can learn more meaningful features by combining the labeled data with the generated pseudo-labels, thereby significantly improving model performance without adding labeled data. Although this framework significantly improves the stability and robustness of the model, quality control of the pseudo-labels and the convergence speed of the student model remain key challenges for future research.

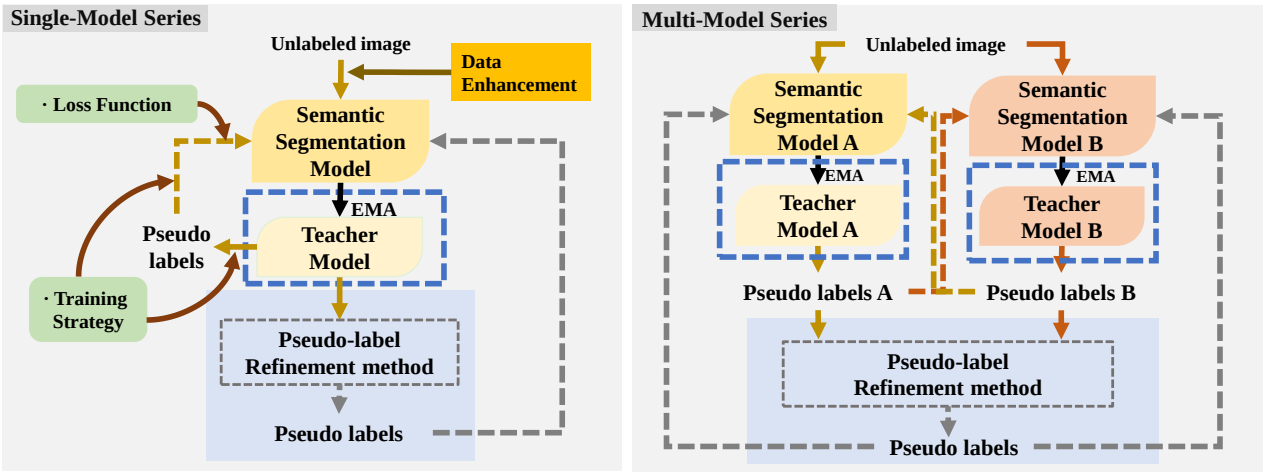


Fig. 4: This illustration shows a graphical representation of the flow of our investigation. Our overview is divided into three main categories: pseudo-label generating models, pseudo-label refinement, data enhancement, and optimization. Within the generation category, we investigated two approaches: single-model and multi-model families (subfigures a and b). In addition, we used various strategies to select or purify pseudo-labels. Finally, optimization methods are also becoming more common (best viewed in colors).

Self-training Iteration. Self-training iteration is a strategy for dynamically updating pseudo-labels, aiming to continuously improve the quality of pseudo-labels and the performance of the model through multiple iterations. GIST and RIST [86] proposed greedy algorithmic strategies (GIST) and iterative self-training-based strategies (RIST) to improve the accuracy of the model by alternating between the real labels and pseudo-labels. The main challenges are how to effectively balance the weights between labeled and pseudo-labeled data, and how to prevent the accumulation of noise in pseudo-labels from affecting the training results. However, this method faces a central problem of noise accumulation in pseudo-labeling [74]. Low-confidence or erroneous pseudo-labels generated by the model can be incorrectly reinforced in subsequent iterations, leading to unstable model performance on under-labeled data. To address this problem, researchers have proposed various improvements such as ST++ [33], a method whose key step is selective retraining. The stability of the generated pseudo-labels during training is utilized in the iterative process to select more reliable unlabeled images, which are filtered by an evolutionary process. Unlike methods that require manually setting pixel filtering confidence thresholds [43], [87], ST++ predicts the performance of pseudo-labels by their gradual change in reliability, thus eliminating the need for pixel-by-pixel confidence selection. CISC-R [88] is introduced in the self-training iteration to labeled image querying to improve pseudo-label accuracy, an innovation that solves the problem of high difficulty in pseudo-label selection in existing methods, but still needs to address how to balance the variability between labeled and unlabeled samples on complex datasets. In addition, PseudoSeg [72] focuses on the redesign of pseudo-labels by introducing a single-stage framework that utilizes calibration fusion techniques to merge labels from two sources (model outputs and class activation maps). This technique has demonstrated a strong potential to improve the quality of pseudo labels through iterative improvement. The iterative nature increases the computation time while balancing the labeled and unlabeled data is crucial to prevent overfitting.

Self-cross Supervision. Self-cross supervision can be used to obtain diverse predictions by different output modules or sub-models supervising each other by exchanging pseudo-labels generated from their predictions, thus providing different perspectives to the model and enhancing the robustness of the learning process. [32] proposed a method called uncertainty-guided self-cross-supervision (USCS) for SSSS, which uses multiple output pseudo-labels from a multiple-input-multiple-output (MIMO) model to supervise each other and employs uncertainty as a guiding message to encourage the model to focus on the high-confidence region of the pseudo-labels (with lower uncertainty) to mitigate the effect of erroneous pseudo-labels in self-cross-supervision, which significantly reduces the number of parameters and computational cost compared to cross-supervision [89]. Significantly reducing the number of parameters and computational cost. Self-cross Supervision provides significant improvements in reducing the parameters and computational overhead of traditional cross-supervision methods. By exchanging different predictions between sub-models, the method improves pseudo-labeling quality. However, it relies on uncertainty scores to guide supervision, increasing the complexity of model design and possibly requiring additional tuning.

Auxiliary Tasks. Auxiliary tasks provide additional supervision by assigning secondary tasks to the network. These secondary tasks typically help improve the segmentation by focusing on specific elements such as boundaries, residuals, or error localization. Earlier, [90] introduced residual networks to enhance the self-training framework. In this method, labeled data is input into an auxiliary residual network to predict the

residuals from the original outcomes. The later proposed Error Localization Network (ELN [91]), on the other hand, primarily aims to assist in error localization. This additional module is trained to detect potentially incorrect pixel points using the images and segmentation results as inputs. The ELN structure contains the main segmentation network (encoder and decoder) and the auxiliary decoder (D_1, D_2, ..., D_K). The main segmentation network is trained by standard cross-entropy loss, while the auxiliary decoder is trained by restricted cross-entropy loss:

$$\mathcal{L}_{aux} = \frac{1}{|D_L|} \sum_{X \in D_L} \sum_{k=1}^{K} \mathbb{1}\left\{L_{ce}\left(P^k, Y\right) > \alpha^k \cdot L_{ce}\left(P, Y\right)\right\} \cdot L_{ce}\left(P^k, Y\right) \quad (4)$$

where $P^k = D^k(\varepsilon(X))$ denotes the segmented prediction of the kth decoder and α^k denotes the scaling hyperparameter used to limit the loss of the kth auxiliary decoder. The performance of the auxiliary decoder will be much worse than the main decoder because it contains various plausible error predictions and uses them as training inputs for ELN, similar to manually creating some error data to train the model. ECS [92] utilized the discrepancy between the primary segmentation network's predictions and the ground truth labels' predictions as a subset of the training images for refinement. EPS++ [93] provides rich boundaries through the saliency map generated by the saliency detection model, which is combined with image-level labeling information for joint training, assisting the model in being trained from pixel-level feedback. Secondary tasks greatly improve the primary segmentation model by providing additional information. Due to the additional complexity of designing these tasks and the increased computational load, a careful balance between auxiliary and primary tasks is needed to prevent overfitting.

Labeled-unlabeled Interactive. This type of approach in which labeled and unlabeled data are processed interactively to improve the overall pseudo-label quality. It is important to highlight that [94] argues handling labeled and unlabeled data independently often leads to the loss of significant prior knowledge obtained from labeled samples. Consequently, they introduced a technique named GuidedMix-Net, which enhances the quality of pseudo-labels by leveraging labeling information to direct the learning process of unlabeled data. In a similar vein, CISC-R [88], as previously discussed, refines pseudo-labels at the pixel level by assessing the pixel similarity between the unlabeled image and the referenced labeled image, subsequently creating the CISC map. Recently, AllSpark [95] leveraged unlabeled data to rebuild the features of labeled data using a transformer architecture, which improves the labeled data and the training model by integrating the entire data stream, thereby addressing low-quality pseudo-labels and minimizing training bias. Labeled-unlabeled interaction, while a powerful strategy, must be carefully balanced between labeled and unlabeled data to prevent the network from over-relying on several labels, especially in the early stages of training.

Pseudo-labels produced by a single model in self-training is often unreliable [96]. Typically, a single model's prediction confidence is used to filter out low-confidence pseudo-labels, which leaves behind high-confidence errors and wastes many low-confidence correct labels [97].

2) Dual-model and Mutual-training:

To address the issues of single-model self-training, where a single model cannot detect and correct its errors, leading to confirmation bias [60] and negatively impacting training and segmentation, mutual-training [119] (illustrated in Fig.4. (b)) is proposed. In this method, two or more models train each other based on their differences, identify their errors, and correct each other. Wang et al. [120] demonstrated that the co-training process can be seen as combined label propagation on two views, offering a viable solution for integrating graph-based and divergence-based semi-supervised methods into a unified framework. Similar methods in semi-supervised learning beyond PL are classified as co-training [121], multi-view constraints [122], and more.

Cross-Pseudo Supervision. Cross-Pseudo Supervision is a two-model technique in which pseudo-labels generated by one network are used to supervise the other network, encouraging mutual refinement and minimizing pseudo-labeling errors. A traditional inter-training method is dual-model cross-supervision. For instance, the CPS [89] approach employs different initialization techniques for two networks, where the pseudo-labels generated by one network are used to supervise the other segmentation network. The later introduced n-CPS [98] extends CPS to n sub-networks. Similarly, a recently proposed work TorchSemiSeg2 [105] uses a cross-pseudo supervision strategy. In addition, there are several related studies for dual models that propose cross-view cross-supervision methods. In their work, [100] introduced a conflict-based cross-consistency (CCVC) technique aimed at compelling two subnets to acquire knowledge from distinct perspectives. To ensure the model extracts more valuable information from conflicting predictions, they introduce a Conflict-based Pseudo-Label (CPL) method that facilitates a stable training process. Specifically, regarding the previous [123], [124], the conflict predictions are further categorized into two types: conflicting and credible and conflicting but untrustworthy predictions. It allows models to solve the problem of confirmation bias by correcting each other's errors through cross-supervision. However, the need to synchronize multiple networks increases computational complexity, and dealing with conflicting pseudo-labels increases design difficulty.

Dynamic Mutual-training. Dynamic mutual training is a technique that initializes two models differently and uses their differences to train each other. DMT [96] is proposed to use models with different initial settings, where one model generates offline pseudo-labels for the other, with differences to identify errors, assesses their divergence, and dynamically adjusts the weight loss from simple to complex [125] throughout the training process to enhance performance. However, during the iterative process, the newly acquired knowledge causes the model to forget previously learned information, leading to the "catastrophic forgetting" issue [126]. The DMT-PLE [127] technique builds upon the pseudo-labels enhancement strategy from the original DMT method. They utilized the pseudo-labels enhancement strategy (PLE) to prevent the model from favoring the most recently learned category, i.e., by using the

TABLE I: An overview of pseudo-labeling techniques in the field of semi-supervised semantic segmentation, for open-source methods we provide links to the code (best viewed in colors).

Category			Method	Publication	Main contributions
Model Structure	Single Model	Self-training Iteration	GIST & RIST [86]	CRV_{21}	Greedy and Randomized Iterative Self-Training
			ST++ [33]	$CVPR_{23}$	Select reliable images for retraining
		Self-cross Sup.	USCS [32]	$ACCV_{22}$	Uncertainty leads self-cross Sup.
		Auxiliary Tasks or Network	[90]	$PRCV_{21}$	Residual correction approach
			ELN [91]	$CVPR_{22}$	Auxiliary Modules:Error localization network, ELN training with constrained loss function
			EPS++ [93]	$TPAMI_{23}$	Explicit pseudo-pixel supervision, Using saliency as pseudo-pixel supervision
		Labeled-unlabeled Interactive	GuidedMix-Net [94]	$AAAI_{22}$	Label information guide unlabeled learning
			AllSpark [95]	$CVPR_{24}$	Unlabeled data reconstruction enhanced Labeled data features; Pure Transformer architecture
	Mutual Model	Cross Pseudo Sup.	CPS [89]	$CVPR_{21}$	Cross-pseudo-supervision
			n-CPS [98]	$ArXiv_{21}$	N networks cross-pseudo supervision
			UCC [99]	$CVPR_{22}$	Uncertainty-guided cross-head co-training
			CCVC [100]	$CVPR_{23}$	Conflict cross-view consistency, Co-training two network branches
		Dynamic Mutual-train	DMT [96]	PR_{22}	Dynamic mutual training
			DMT-PLE [77]	$Access_{22}$	Pseudo-label enhancement strategy
Pseudo-label Refinement	Label Update	Pseudo-label Corrections	[78]	TIP_{21}	Graph-based noise labels correction
			CARD [79]	$IJCAI_{22}$	Semantic linking to correct noisy labels
			ThreeStageSelftraining [77]	TIP_{22}	Multi-task count and update pseudo-labels
			CISC-R [88]	$TPAMI_{23}$	Labeled images correct inaccurate pseudo-labels
			LogicDiag [101]	$ICCV_{23}$	Logic rule abstraction semantics for optimization
			CorrMatch [71]	$CVPR_{24}$	Modeled feature maps for correction
		De-biasing	DARS [102]	$ICCV_{21}$	Redistribute biased pseudo-labels, aligned with the true distribution
			DST [103]	$NeurIPS_{22}$	De-biased self-training
	Filter-only	Confidence Filtering	C3-SemiSeg [104]	$ICCV_{21}$	Dynamic Confidence Region Selection Strategy
			CAFS [97]	$ArXiv_{23}$	Adaptive classification confidence thresholds
			TorchSemiSeg2 [105]	$ICME_{23}$	Local Pseudo Labels Filtering Module
		Confidence Refinement	PGCL [106]	$WACV_{23}$	Network Pruning Refinement Confidence Score
		Assisted Net Filtering	GTA-Seg [107]	$NeurIPS_{22}$	Assistants Teacher sift useful information
Data Enhancement	Classic Data Enhancement		CutMix [108]	$ICCV_{19}$	Cut and paste patches in the training images
			ClassMix [109]	$WACV_{21}$	Mixing of unlabeled samples to generate enhanced images and artificial labels
			Complexmix [110]	$ICIP_{21}$	Mask-based data enhancement combining aspects of CutMix and ClassMix
			SimpleBaseline [111]	$ICCV_{21}$	Effective use of strongly augmented baselines in pseudo-labels learning
			AugSeg [112]	$CVPR_{23}$	Simplified intensity-based enhancement, selecting a random number of respective data enhancements
			iMAS [113]	$CVPR_{23}$	Dynamically adjusted data augmentation and model adaptive supervision for each instance
	Geometric Data Enhancement		MT-PhTPS [114]	MVA_{21}	One-way consistency and nonlinear geometric and photometric perturbations
			ADS-SemiSeg [115]	$TCSVT_{22}$	Context-friendly geometric warping for unlabeled image perturbations
			semisup-seg-efficient [116]	$Sensor_{22}$	Competing perturbations consisting of geometric distortions and photometric dithering
Optimization Improvement	Loss Function	Cross-Entropy Loss Improvement	Learning Pseudo-Label [117]	PR_{22}	Class-aware Cross-entropy loss
		Weights Guide	PS-MT	$CVPR_{22}$	Confidence-weighted Cross-entropy (Conf-CE)
			[34]	$CVPR_{20}$	Improvement of pseudo-label boundary accuracy
		Additional Ancillary Loss	CorrMatch [71]	$CVPR_{24}$	Combining cross-entropy loss with consistency
	Training Strategy	Curriculum Learning	Learning Pseudo-Label [117]	PR_{22}	Progressive cross-training gradually introduces high-quality predictions as additional sup.
			PGCL [106]	$WACV_{23}$	Step-by-step learning based on the confidence scores predicted by the model
			ESL [69]	$ICCV_{23}$	Dynamically maintaining dominant categories (soft pseudo-label)
		Two-branch Exploration	CPCL [118]	TIP_{23}	Conservative & progressive exploration

pseudo-labels generated by the previous stage of the model to refine those produced by the current model. It is an effective method though to combine the strengths of both models while minimizing their weaknesses. Due to its two-model structure, this approach requires more computational power.

This section delves into the nuances of both single-model and dual-model methods for SSSS. Single-model, while simpler, faces challenges like noise accumulation in pseudo-labels and computational inefficiency. Dual-model, such as cross-pseudo supervision and dynamic mutual training, provide more robust training but come at the cost of increased complexity. Together, these methods advance the state of semi-supervised learning by providing innovative solutions to the inherent challenges of working with both labeled and unlabeled data.

C. Pseudo-label Refinement Methods

Noise is an inevitable aspect of PL. Introducing erroneous predictions or faulty labels during training results in error buildup, which obstructs proper guidance for further learning and negatively affects the segmentation model's training results. To address this issue, a range of techniques have been introduced in related research. In this subsection, we will examine these techniques from two viewpoints: the first involves filtering-only methods, used to remove unreliable pseudo-labels, and the second involves pseudo-label updating methods, aimed at fixing or correcting pseudo-labels.

1) Filtering-only Methods: Filtering-only methods focus on filtering out noisy pseudo-labels based on confidence to improve segmentation results. A threshold is set to filter out unreliable predictions through fixed or dynamic methods, sometimes with the help of an auxiliary network.

Fixed Threshold Filtering. Fixed threshold filtering uses a static, predefined threshold to determine the reliability of pseudo labels. During training, only pseudo-labels with a confidence level higher than the threshold are used to avoid incorporating noise into the model. FixMatch [74] is an image classification approach that relies on pseudo-labels. During the generation of pseudo-labels, if the confidence in image classification exceeds a fixed threshold, the loss for that image is calculated. In the realm of semantic segmentation, some techniques employ straightforward threshold filtering. CAC [43] utilized a fixed threshold to create pseudo-labels. U2PL [64] applied a fixed entropy value as the filtering criterion for the prediction of each pixel. DST-CBC [76] set the threshold based on the proportion and overall confidence distribution of a specific class of pixels and progressively increased the proportion of pixels in the pseudo-labels throughout training. Similarly, DGCL [128] adjusted this proportion based on the entropy of each pixel's prediction result. While fixed thresholds are simple to use, their fixity may discard potentially useful low-confidence labels or retain some high-confidence but incorrect labels, and the inability to adapt to model predictions that evolve is also a drawback.

Dynamic Confidence Filtering. Unlike the fixed threshold approaches, dynamic confidence filtering adjusts the confidence threshold during training based on the changing performance of the model. This allows more flexibility in retaining pseudo-labels that initially have low confidence, but whose confidence improves over time. C3-SemiSeg [104] employs a dynamic selection strategy for confidence regions to concentrate on high-confidence areas during loss calculation. Furthermore, it incorporates cross-set contrast learning to enhance feature representation. [75] suggests that using a thresholding method based on categories might be more effective. To tackle the issue where existing high-reliability pseudo-labels discard significant information, [97] introduced Adaptive Class-by-Class Confidence Thresholding (ACT), which reduces the reliance on calibration scores to dynamically modify the reliability confidence thresholds. Given that much of the prior research evaluated pseudo-labels primarily based on confidence thresholds, the initial issue of confidence ambiguity may significantly hinder subsequent updates. Recently, [106] proposed PGCL, which addresses the challenge of unclear confidence scores in network pruning by incrementally training the network from simple to complex examples using a coarse strategy. As noted in FreeMatch [129], using fixed thresholds that are either too stringent or too lenient can negatively impact model convergence. Corrmatch [71] introduces a dynamic global thresholding strategy that adapts to the training process. Dynamic Confidence Filtering avoids the negative impacts of too strict or loose fixed thresholds, especially in the later stages of training, it can better capture the details and effectively deal with the confidence changes in complex scenes or different categories. However, the threshold needs to be adjusted according to different categories or scenes, which increases the computational complexity. It may introduce new instability, which may affect the convergence of the model if it is not properly adjusted.

Assistant Network Filtering. Auxiliary network filtering relies on an auxiliary network to assist in filtering. This auxiliary (or assistant) network helps the primary network to pass only reliable pseudo-labels, filtering out noisy or unreliable pseudo-labels. Moreover, certain techniques employed auxiliary frameworks for filtering tasks. For instance, GTA-Seg [107] employed an auxiliary framework called the gentle Teaching Assistant. The GTA acquires knowledge directly from the pseudo-labels created by the teacher's network, and solely filtered, beneficial information is transmitted to the student's network to aid in overseeing the student's network training. The auxiliary network helps to provide higher-quality training data for the main network, which further improves the quality of the pseudo-labels. However, the additional network structure increases the training time and consumption of computational resources. If the auxiliary network itself is not of high quality, new noise or bias may be introduced.

2) Pseudo-label Update Methods: Unlike filter-only methods that discard unreliable labels, pseudo-label update methods focus on improving the quality of pseudo-labels by proactively correcting and refining them during the training process. Label correction and debiasing techniques can ensure that updated pseudo-labels provide more accurate guidance and help improve training stability and efficiency.

Label Corrections. Label correction methods aim to detect and correct noisy pseudo-labels by comparing them with other sources of information (typically using secondary networks,

graph-based methods, or external knowledge). The goal is to reduce noise and provide more accurate pseudo-labeling for training. Pseudo-labels contain noise. Several approaches attempt to identify and rectify these errors by incorporating extra indicative categories [92], residual errors [90], or deploying additional networks [35]. Some researchers have approached this task as learning from noisy pixels. [78] proposed a framework for detecting and correcting labels using a graph-based approach, which employs a graph attention network supervised by clean labels to mitigate noise. CARD [79] introduced a relational network that is independent of categories, aiming to correct labels by leveraging dependable semantic relationships between image features. DMT-PLE [77] offers a strategy to enhance pseudo-labels by updating them with prior knowledge accumulated from earlier iterations. Drawing inspiration from ST++ [33], CISC-R [88] employed a query-based image selection technique that considers inter-class feature variations and the challenges of noise correction at the initial stages of training. LOGICDIAG [101] introduced a novel approach by using symbolic knowledge and logical rules to abstract semantic concepts, aiming to identify and rectify incorrect labels. Recently, several have been introduced to focus on specific region corrections. TorchSemiSeg2 [105] employed a discriminator to evaluate the trustworthiness of region-level labels and perform region corrections. CorrMatch [71] utilized feature maps for correction. In contrast to these approaches, some enhancement techniques are initiated from the training phase. [77] proposed a Three-Stage Self-Training method that generated initial pseudo-labels for unlabelled data through a self-training process, ensuring segmentation consistency across multiple tasks. These approaches detect and correct noise, ensuring more accurate training data and reducing error propagation. The effectiveness is highly dependent on the quality of external information, and if the correction strategy is not appropriate, new biases can be introduced instead.

De-biasing. The focus of De-biasing is to reduce bias due to noise or incorrect pseudo-labels during training. Bias accumulates throughout iterations, resulting in errors that are reinforced over time. Such methods aim to correct this situation by redistributing biased pseudo-labels to be consistent with the real data distribution. Training bias originates both from the network itself and from the improper training of potentially incorrect pseudo-labels, which accumulate errors over iterations. [102] introduced Distributed Alignment and Random Sampling (DARS), a straightforward and effective technique to redistribute biased pseudo-labels, align them with the ground truth, and mitigate the impact of noisy labels on training. Aligning their distribution with the true distribution enhances SSSS. To further reduce bias, DST [103] proposed debiased self-training. This method's core principle is that two parameter-independent classifier heads decouple the generation and utilization of pseudo-labels, ensuring that only clean labels are used for training. By re-removing the bias in training and assigning the distribution of pseudo-labels, these solutions ensure that they are aligned with the distribution of the real data, which improves the model's generalization ability and accuracy. Strategies need to be precisely designed and may not be effective in removing bias in pseudo-labels if not properly implemented. Meanwhile, additional debiasing steps may increase the training complexity.

In this section, we discuss pseudo-label optimization from two aspects: filtering and updating. Filtering methods focus on improving model robustness by removing low-confidence labels and preventing noise from entering the training process. In contrast, pseudo-label update methods aim to continuously improve the accuracy of pseudo-labels and ensure that the model learns more reliable pseudo-labels. Each of these two types of methods has its focus, with the filtering methods effectively reducing the noise interference in the pseudo-labels, while the updating methods further optimize the model's learning process by improving the quality of the pseudo-labels.

D. Data Enhancement

Data enhancement plays a crucial role in PL for SSSS in the context of limited labeled data for maximizing model performance using both labeled and unlabeled data. These methods aim to artificially increase the diversity of the training data by performing various transformations, enhancements, or perturbations on the existing datasets. This subsection explores two broad classes of data augmentation techniques: Classic-enhancement Methods and Geometric-enhancement Methods, examining their evolution, advantages, and limitations.

1) Classic-enhancement Methods: Classical augmentation methods are effective in improving the generalization ability of models mainly by creating more diverse training data.

Yuan et al. [111] introduced a straightforward yet highly effective framework as a baseline for employing a robust family of data augmentation techniques. They emphasized that attention to detail is crucial and that a straightforward combination of design and training methods can greatly enhance segmentation performance. To enhance model generalization, [108] introduced the CutMix augmentation technique in their study. This classical method, frequently cited in later research, involves cutting and pasting patches within training images to effectively leverage the regularization effect of training pixels and the loss of preserved regions. ClassMix [109] created augmented data by combining two unlabelled samples to produce synthetic images and pseudo labels. Following this, [110] introduced ComplexMix, a novel mask-based data augmentation technique that integrates the CutMix and ClassMix strategies. AugSeg [112] randomly chooses various data augmentation techniques to adaptively enhance unlabelled samples, guided by the model's estimated confidence for each sample. They further contend that most current research overlooks the variability among unlabelled instances and the challenges in training. To address this, they recently proposed instance-specific and model-adaptive supervision (iMAS) [113]. Their modeling concepts are designed to pay more attention to the unique attributes of each data point, aiming to target and improve model performance. The simplicity and adaptability of classical enhancement methods make them easy to implement in a variety of tasks. However while these methods can improve model performance, they typically focus on pixel-level transformations and may not fully capture more complex changes in the data distribution.

2) Geometric-enhancement Methods: Geometric enhancement takes a different approach and usually modifies the spatial configuration of the image without changing the pixel values. These techniques are particularly useful for introducing spatial diversity in the training data, thus making the model more robust to geometric distortions.

Cao et al. [115] introduced a context-aware micro-aggregable warping technique, which is an unsupervised strategy primarily aimed at perturbing unlabeled images. M3L [116] introduced a robust perturbation model, which incorporates geometric distortions and photometric variations to achieve consistent predictions on unlabelled, perturbed inputs. Following this, [114] advanced the linear method for addressing geometric enhancement by introducing a baseline called MR-PhTPS, which relies on nonlinear geometric and photometric perturbations. SemiFL [130] grounded in data enhancement integrated semi-supervised federated learning. Geometric enhancement specializes in dealing with spatial data variations, making models more resilient to distortions and geometric changes, and their main advantage is their ability to capture complex spatial structures. Due to the complexity of geometric transformations, they may require more computational resources and their impact on tasks less sensitive to spatial alignment may be limited.

This subsection explores the key role of data enhancement in improving the performance of PL when labeled data is limited in SSSS. Data enhancement aims to increase data diversity by perturbing the data. There are two categories: Classical-enhancement and Geometric-enhancement. Classical enhancement is widely used due to its simplicity, and geometric enhancement improves the robustness of the model to complex structures by dealing with spatial transformations of the image. However, classical solutions tend to favor pixel-level transformations and may not be able to capture more complex variations in the data distribution, while geometric enhancement, while excelling in spatial variations, may face higher computational costs. In practice, a reasonable combination of these enhancement methods is expected to further optimize the efficiency and performance of the model in utilizing the data in the case of annotation scarcity.

E. Optimization Improvement Methods

Enhancing optimization methods is key to boosting PL performance. In this part, we will concentrate on two major components: Loss Function and Training Strategy, each playing a vital role in tackling issues with unlabeled data. Loss functions are often adjusted or crafted to improve confidence, minimize noise, and enhance the learning process. Concurrently, the training strategy directs the model in making the most out of both labeled and unlabeled data using approaches like curriculum learning and dual-branch exploration.

1) Loss Function: Loss functions are a critical component in any deep learning model, dictating how well the model learns from data and adjusts its parameters. In SSSS research, the standard cross-entropy (CE) loss remains widely used, but several modifications have been introduced to improve the quality of pseudo-labels, enhance training, and address challenges such as class imbalance and label noise.

Cross-Entropy Loss Improvement. Due to challenges such as category imbalance and pseudo-labeling noise, traditional CE may not be sufficient, and several modifications for CE have been proposed to improve pseudo-label generation and quality. [117] argued that it is very challenging to summarise the common features of classes and classify them based on a small number of samples, so the CE loss should be uniquely designed based on the SSL properties, and a category-aware cross-entropy (CCE) was proposed to reduce the class candidates by introducing image-level labels:

$$\mathcal{L}_{\mathrm{CCE}}(\mathbf{x},\mathbf{y},\mathbf{c}) = -\frac{1}{P}\sum_{p=1}^{P}\sum_{i=1}^{C}\mathbf{y}_{i,p}\log\frac{e^{\mathbf{x}_{i,p}}}{\sum_{i=1}^{C}\mathbf{c}_i e^{\mathbf{x}_{i,p}}} \tag{5}$$

where $\mathbf{x}_{i,p}$ is the pixel-level network output of class i at position p, $\mathbf{y}_{i,p}$ are the corresponding segmentation labels, and the CCE normalizes the prediction $\mathbf{x}$ only for those classes that appear in the current image ($c_i = 1$). CCE simplifies the generation of pseudo-labels compared to conventional CE. PCT incrementally incorporates high-quality predictions as supplementary guidance for network training. Also from the category perspective, [59] addresses the category imbalance problem of pseudo labels by proposing a balanced loss function to reduce the dominance of frequent classes. The Cross-Entropy Loss Improvement addresses the core limitations of traditional CE loss, such as category imbalance and sparse data representation. By focusing on the categories present in each image, CCE reduces the noise introduced by irrelevant categories. However, these improvements require careful tuning and may be sensitive to specific dataset features, which limits their generalizability across different tasks.

Weights Guide. Weights Guide introduces a confidence-based weighting mechanism that adjusts the importance of individual predictions based on confidence scores to deal with noisy labels and improve the stability of training. Confidence threshold-based selection [131] is also a very commonly used perspective, [132] which reduces noisy labels by weighting predictions that encourage a higher level of confidence. of the predictions. PS-MT [133] adopts a stricter confidence-weighted cross entropy (Conf-CE) as the loss to alleviate the problem that CE loss training tends to lead to excessive prediction errors. The specific function is represented as follows:

$$\ell_{con}(D_U,\theta^s) = \frac{1}{|D_U||\Omega|}\sum_{\mathbf{x}\in D_U}\sum_{\omega\in\Omega} c(\omega)\ell(\tilde{\mathbf{y}}(\omega), p_{\theta^s}(\mathbf{x})(\omega)) \tag{6}$$

where $\ell(.)$ denotes the CE loss, ω denotes the pixel address of the output lattice Ω of the segmentation map, $\tilde{\mathbf{y}}(\omega) \in \{0,1\}^Y$ is the prediction of the teacher model from ω, i.e., the pseudo-label, $p_{\theta^s}(\mathbf{x})(\omega) \in [0,1]^Y$ is the segmentation prediction of the student model, and $c(\omega) \in [0,1]$ denotes the prediction confidence level of the teacher model. Conf-CE can be bounded to the high confidence segmentation result of the region computation. While [134] uses uncertainty-weighted loss to adjust their contribution in training based on the uncertainty of the pseudo-labels. [34] pays special attention to improving pseudo labels in the border region labels accuracy. Weights Guide greatly improves the model's ability to filter

out noise in pseudo-labels. By focusing on high-confidence predictions, the model avoids overfitting noisy data, resulting in more robust learning. However, determining the appropriate confidence threshold and weighting scheme can be challenging and requires extensive experimentation to strike a balance between reducing noise and retaining useful information.

Additional Ancillary Loss. In addition to CE, additional losses such as consistency and contrast learning losses are introduced to improve the stability and feature representation of the model, and these auxiliary losses help to ensure that the model behaves consistently across different enhancement processes and optimize the pseudo-labels based on feature similarity. Consistency loss [84], [135] focuses on enhancing the identical response of the model to the input data, thus effectively improving the stability of the model in real-world applications again. [136]–[138]combines the cross-entropy loss with the consistency loss to improve the reliability and accuracy of pseudoclabels. Some studies incorporate consistency into the calculation of CE loss, such as those mentioned in Corrmatch [71]:

$$\mathcal{L}_u = \frac{1}{N}\sum_{i}^{N} \ell(\mathcal{F}(x_i^s), \mathcal{F}(x_i^w)) \odot \mathcal{M}_i \tag{7}$$

where x_i^w and x_i^s denote the unlabelled image and its weakly enhanced and strongly enhanced versions, respectively. Corrmatch treats the prediction of x_i^w as a pseudo-label for x_i^s and encourages the output to be consistent under both weakly and strongly enhanced inputs. U2PL [64] and DGCL [139] combine contrastive learning loss to optimize pseudo-labels by feature similarity between samples, which guides the model to learn more accurate and robust feature representations. Auxiliary losses are critical to making SSL models more adaptable to changes in the input data. More robust pseudo-label refinement can be ensured by enforcing consistency and encouraging feature alignment. However, the trade-off between computational complexity and performance gains needs to be considered, and additional losses can increase training time and resource requirements.

2) Training Strategy: In PL for SSSS, a reasonable training strategy is crucial to effectively utilize both labeled and unlabeled data. This section explores two representative training strategies: Curriculum Learning and the Two-branch Exploration. Curriculum learning guides the model from simple to complex by gradually increasing the task difficulty, while the Two-branch Exploration strategy fully exploits unlabeled data through co-evolutionary exploration.

Curriculum Learning. The PL relies heavily on the reliability of the obtained prediction labels, but unlabelled data is not equally reliable in SSL training, and the curriculum learning strategy, which actively selects credible samples based on the model's current capabilities and gradually introduces more complex samples, can to some extent solve this problem. [106] gradually learned based on the confidence scores of the model predictions, somewhat suppressing the introduction of noisy information. ESL [69] improved the overall segmentation performance by dynamically maintaining the dominant category (soft pseudo-labels) i.e., more confident prediction for each pixel, and gradually transitioning to highly uncertain samples as the training progresses. [117] then gradually introduce high-quality predictions as additional supervision for network training through progressive cross-training. Curriculum Learning provides an elegant solution to the noise-pseudo-labeling problem by starting with simpler, high-confidence data, allowing the model to build a solid foundation before dealing with more complex data. However, an incremental approach may slow down training because of the need to iteratively assess the confidence of the predictions and adjust the difficulty of introducing samples. In addition, determining the optimal progression of sample complexity can be tricky, especially in highly heterogeneous datasets.

Two-branch Exploration. This is an innovative training strategy aimed at balancing exploration and stability in SSSS, with the core idea of exploring the full utilization of unlabeled data using two parallel prediction networks. CPCL [118] (Conservative-Progressive Collaborative Learning) enables two branches to operate under different supervised paradigms: cross-supervision and joint-supervision. The conservative branch seeks commonalities between predictions, ensuring reliable model updates based on the most certain labeled predictions. The progressive branch embraces divergence by utilizing the full set of pseudo-labels, encouraging the model to explore potential patterns in the data. Like other multi-branching approaches (Dual Student [134] and Cross Pseudo Supervision [89]) also explores multistream solutions. The unique feature of the two-branch exploration strategy is that it emphasizes the collaboration and distinction between conservative and progressive branches, thus achieving a finer balance between stabilization and exploration. However, the coordinated two branches need to be carefully designed to ensure that they complement each other effectively.

This section provides a summary exploring two components for improving the performance of PL techniques in SSSS: Loss Functions and Training Strategies. Undeniably, the loss function is the primary mechanism for tuning model parameters and refining it is key to reducing labeling noise and addressing the challenges inherent in using unlabeled data. Standard CE loss, while effective, is often insufficient in SSL, so advanced loss has been developed to mitigate class imbalance and improve label quality. Meanwhile, training strategies such as curriculum learning adapt the training process to predictions with different confidence levels, thus gradually improving model performance, and two-branch exploration guides the effective utilization of labeled and unlabeled data. Together, these approaches enhance the robustness and stability of the training process, addressing the main challenges associated with noisy and unbalanced data, while making a trade-off between complexity and performance.

F. Hybrid Techniques for Pseudo-Labelling Combined with Other Semi-supervised Methods

After an in-depth discussion of PL in SSSS, this section focuses on an innovative approach that further enhances performance by combining PL with other state-of-the-art techniques. Those methods not only retain the advantages of the PL in

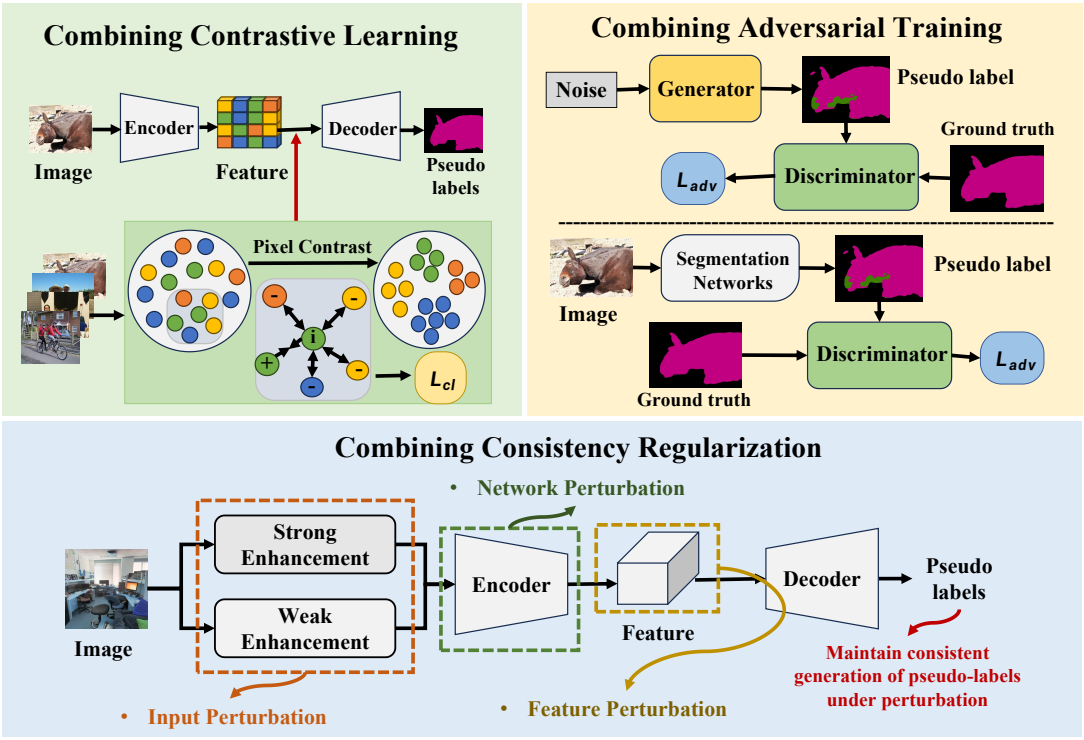


Fig. 5: This illustrated solution details a hybrid approach to PL combining other SSSS advances, including consistent regularization (CR), comparative learning (CL), and adversarial training (GAN), notably adversarial training incorporates two architectures differentiated by whether or not they contain Generators(best viewed in colors).

exploiting unlabeled data, but also enhance the generalization ability and robustness of the model by integrating strategies such as Consistent Regularization (CR), Comparative Learning (CL), and Adversarial Training (AT). We show a detailed illustration of the hybrid technique in Fig.5.

1) Combining Consistency-Regularization Methods:

CR [140] is typically implemented by introducing perturbations at various stages of the training process to train models that produce consistent predictions. The perturbations used in existing studies include input perturbation [36], feature perturbation [141]–[143], network perturbation [35], [144], and hybrid perturbation [145], [146]. Currently, some research in the SSL domain has focused on integrating CR with the PL approach [147]. These studies have demonstrated that combining the two techniques enhances performance, indicating the potential for further improving accuracy in the SSSS domain, particularly for images with limited labeled data. [135] initially developed FixMatch to streamline the SSL process created high-confidence pseudo-labels and integrated them with consistency regularization to guarantee stable prediction outcomes for augmented unlabeled images.

Image and Feature Dual Perturbation. Given that many earlier studies utilized a single type of perturbation and obtained promising outcomes, there have been efforts to integrate different perturbations. A common research direction involves combining image perturbation with feature perturbation. [148] innovatively introduced a novel framework named the Auxiliary Mean Teacher Model, which is grounded in a consistency regularization method. In this framework, pseudo-labels produced by one mean teacher guide another student network, facilitating mutual knowledge distillation between the two branches. CFCG [138] employs image-level and feature-level perturbations and cross-fertilization supervision (CFS) mechanisms to extend the feature distribution. Given that certain SSL methods like SwAV [149] and ReMixMatch [150] utilize multi-branch perturbations to enforce consistency, UniMatch [124] similarly aimed to merge image and feature perturbations and revisited the weak-to-strong consistency framework introduced by FixMatch [135] in SSL and enhanced it by creating a straightforward and effective consistency framework. This framework introduces a dual-flow perturbation approach, expanding the perturbation space by separating image and feature perturbations into two distinct flows.

Resolution of class imbalances. Given that SSSS frequently deals with limited or unbalanced datasets, it often underperforms in certain categories and faces significant category bias issues, such as the long-tailed label distribution seen in the tail categories of the CityScapes dataset. However, many current methods largely overlook this issue and treat all categories the same. The integration of consistent regularization and PL offers a novel approach to addressing class imbalance by enhancing the model's generalization to a few classes and low-confidence predictions. [179] introduced a novel framework, AEL, which adaptively adjusts the training process for well-performing and poorly-performing categories. It dynamically monitors category performance using a confidence base during

training, thereby shifting the training focus towards the underperforming categories. Subsequently, USRN [180] proposed an unbiased subclass regularization network to address category imbalance by learning unbiased category divisions within a balanced subclass distribution using K-means clustering [181].

Cross-View Learning. [99] proposed an uncertainty-guided cross-head co-training (UCC) framework that incorporates both weak and strong augmentations within a shared encoder. This approach seamlessly merges the benefits of CR and PL, enabling the learning of a more compact feature representation from two distinct perspectives. PseudoSeg [72] employed a single-stage consistency framework that effectively combined robust data augmentations from various perspectives and created pseudo-labels for consistency training. CCVC [100] introduced the cross-view consistency (CVC) strategy, which promotes the two subnets to extract distinct features from the same input by incorporating a feature difference loss.

The integration of CR with PL enhances the learning process of SSSS. CR ensures that the predictions remain consistent when perturbed due to its properties, which minimizes the sensitivity of the PL model to changes, resulting in a more stable pseudo-label generation process. Additionally, CR addresses category imbalance and improves model performance on underrepresented categories, which further complements PL. In situations where pseudo-labels may be noisy or of low confidence, CR helps normalize and stabilize the learning process, ensuring that the model remains robust even in the face of challenging datasets or imbalanced categories.

2) Combining Contrastive Learning Methods:

CL [182] aims to enhance semantic segmentation by learning more effective representations in the embedding space. This is achieved by drawing positive samples closer and pushing negative samples further apart, thereby enabling the model to discern pixel similarities and differences better. Given the task's specificity, the common approach involves augmenting query samples with images modified through data perturbation. ReCo [183] initially introduced the application of CL methods to SSL tasks.

Currently, some approaches integrate CL with PL methods. [64] introduced an enhanced technique, U2PL, focusing on unreliable pixels by placing them into a negative sample storage queue to maximize the use of unlabeled data. In a recent study, DGCL [139] introduced Density Guided Contrastive Learning, which aims to move anchor features in sparse areas closer to cluster centers represented by density-positive keys. ELN [91] disregards label noise during the training phase, thereby attaining robustness against erroneous pseudo-labels and can be seamlessly integrated with CL. Other research integrates CR and CL to enhance PL. C3-SemiSeg [104] leveraged CR for feature alignment under perturbations, proposes a new cross-set region-level data augmentation approach and incorporates cross-set CL to boost feature representation.

CL addresses the limitation of confirmation bias in PL by encouraging the model to focus on the relational structure of the data, which enhances the model's ability to distinguish between categories by pushing dissimilar pixels out of the way and bringing similar pixels closer together. Not only does this make the representation more robust, but by organizing the feature space and ensuring better separability between categories, CL complements PL, ultimately enabling more reliable training and better generalization, especially in complex or high-dimensional segmentation tasks.

3) Combining Adversarial Training Methods:

Combining with Adversarial Training (AT) usually generates pseudo-labels through the use of Adversarial Generative Models (GAN) [184], but involves huge training challenges and high inference costs. Due to the complexity of the generation task, these methods are typically restricted to specific application areas, such as medical image segmentation for particular families or diseases. Earlier techniques [37] introduced additional supervision via a generator. Given the training difficulty and computational demands, more recent studies [35], [185], [186] have concentrated on segmentation networks that incorporate only a discriminator. In this network

TABLE II: A summary of common application scenarios of pseudo-label approaches covered in this review, for open-source methods we provide links to the code (best viewed in colors).

Category	Method	Publish	Dataset(s)	Main contributions
Medical Image Segmentation	ATSO [50]	$CVPR_{21}$	[151],etc.	Asynchronous teacher-student optimization algorithm
	SLC-Net [61]	$MICCAI_{22}$	[152], [153]	Dual network cross-model pseudo-supervised framework
	Compete to Win [73]	TMI_{23}	[151], [152], [154]	New competition winning approach, boundary awareness enhancement module
	EMSSL [62]	$ArXiv_{23}$	[155]–[157]	Variational approach to generalize Bayesian pseudo-labels
	BCP [158]	$CVPR_{23}$	[151], [152], [159]	Mean Teacher structure combined with two-way copy-paste supervision
	EPL SemiDG [160]	$AAAI_{22}$	[161], [162]	Confidence-aware cross-pseudo-supervision incorporating∼consistency regularization
	ACTS [163]	TMI_{22}	[159], [164], [165]	Dynamic convolution, adversarial self-integrating network combining consistency regularization and adversarial learning
	S5CL [166]	$MICCAI_{22}$	[167], [168]	Single-stage self-supervised pretext task, model fine-tuning
	CLCC-semi [169]	$ISBI_{22}$	[165], [170]	Cross-level comparison learning and consistency constraints
	CDCL [171]	$CVPR_{22}$	[172], [173]	Cross-patch dense contrast learning framework
Remote Sensing Image Segmentation	[51]	$Elec._{23}$	Vaihingen [174]	Dual network combination: UNet and DeepLabV3
	[175]	$Entropy_{23}$	[63],etc.	Double cross entropy consistency, channel attention mechanism
	ICNet [64]	$GRSL_{21}$	[176]	Iterative contrast network incorporating contrast learning
	MS4D-Net [177]	$Remote\ Sensing_{23}$	[178]	End-to-end multi-tasking Siamese networks incorporating consistency regularization

structure, the segmentation network acts as a generator, while the discriminator assesses the authenticity of the images.

Cao et al. [115] introduced an adversarial two-student framework named ADS-SemiSeg, which enhances the "Mean Teacher" approach in two primary aspects. Firstly, the two-student model is trained independently while incorporating stability constraints to leverage model diversity. Secondly, an adversarial training strategy is employed for both students, along with a discriminator to identify reliable pseudo-labels from the unlabeled data for self-training. Referring specifically to s4GAN [70], the discriminator functions as the adversarial component in each branch.

The combination of AT and PL provides a complementary approach to enhance SSSS by addressing the different weaknesses of each method. While PL provides a mechanism for utilizing large amounts of unlabeled data, it often suffers from noisy labels. AT improves network robustness by incorporating a discriminator that filters out unreliable pseudo-labeling, thereby mitigating noise and increasing the reliability of the training data. This combination compensates for each other's limitations: while PL extends the dataset with automatically generated labels, AT ensures that only the most trustworthy pseudo-labels are retained for further learning, leading to a more stable and efficient training process, especially in challenging domains where labeled data is scarce.

III. Pseudo-label Methods in Other Areas

Pseudo-labeling is extensively utilized in semantic segmentation due to its straightforwardness and efficiency. The techniques discussed in the previous section primarily target natural image segmentation. However, advancing research and application of image segmentation across various fields is undeniably crucial. This section concentrates on pseudo-label methods applied to specific areas, such as medical imaging and remote sensing image segmentation. Table II provides a summary of the PL methods discussed in this section.

A. Medical Image Segmentation

The process of segmenting medical images, which involves identifying the pixels corresponding to organs or lesions in CT or MRI scans, is highly challenging in medical image analysis due to the limited availability of sufficient labels. Numerous studies have proposed using PL, demonstrating promising outcomes when applied to specific medical datasets.

Initially, [50] identified a flaw in the model known as lazy mimicking, where the model tends to maintain its previous predictions and resist change. They proposed a new method called Asynchronous Teacher-Student Optimization (ATSO) to solve the inert problem. In contrast, EMSSL [62] improves the interpretability of the model to a certain extent by connecting the original method to the Expectation Maximization (EM) algorithm [187] and providing a comprehensive generalization of the pseudo-labels within a Bayesian framework. However, a frequent issue in medical image segmentation is the uneven distribution of labeled and unlabeled data. Prior research that handled these two data segments separately or inconsistently might overlook the valuable insights from the labeled data. BCP [158] suggested a simple architecture to address this issue by integrating labeled and unlabeled data bi-directionally. Moreover, several studies have implemented the dual-model concept to enhance PL for medical image segmentation. [61] introduced an innovative cross-model pseudo-supervision framework, SLC-Net, which leverages shape awareness and local context constraints to produce anatomically accurate predictions. [73] created high-quality pseudo-labels by comparing multiple confidence graphs generated by different networks and selecting the result with the highest confidence.

The scarcity of annotations and the abundance of unlabeled data are common challenges in computational pathology segmentation. SSL is a solution to this problem, but a single approach often fails to achieve satisfactory results, and many reports have PL hybrid techniques that have been extensively explored and investigated. [166] integrated PL with CL to create a unified medical segmentation framework, S5CL, that combines the two training phases of self-supervision and model fine-tuning into one. In addition, some studies utilized the integration of CL and CR for SSSS in medical images. [169] proposed a method called CLCC, which merges cross-level CL with CR constraints to improve the representation of local features in the medical field. Following this, [171] introduced CDCL, a dense CL framework based on cross-patch analysis, aimed at segmenting cell nuclei in pathological images. To improve the quality of pseudo-labeling in medical segmentation, [160] proposed an innovative confidence-aware cross-pseudo-supervised algorithm, EPL-SemiDG, which combines PL and CR techniques. ASE-Net [163] have explored the combination with adversarial learning on top of this, leading to the proposal of a novel adversarial consistency self-integrating network.

B. Remote Sensing Image Segmentation

Annotating high-resolution remote-sensing satellite images is a time-consuming and labor-intensive task. This constraint impacts the effectiveness of segmentation models. Some studies suggest employing PL techniques based on SSL to mitigate this challenge. These approaches are designed to aid in the segmentation of remote-sensing images.

Similarly, the availability of high-quality labeled images for remote sensing is quite limited, just as it is in medical image segmentation. Several reports have introduced semi-supervised methods to train remote sensing image segmentation models, achieving remarkable results. Wang et al. [64] introduced ICNet for remotely sensed images, which utilizes the CL technique to obtain more potential information in remotely remote sensing images gradually. In [177], PL and CR are integrated to develop MS4D-Net, an end-to-end framework for assessing post-disaster building damage in remote sensing image segmentation. This framework employs a multi-task Siamese network to enhance damage classification outcomes by utilizing building extraction results.

Li et al. [51] introduced a method that employs two networks (UNet [200] and DeepLabV3 [201]) to predict labels for the same set of unlabeled samples. They then incorporate pseudo labels with high predictive consistency into the training set to enhance the accuracy of semantic segmentation when

TABLE III: A compilation of datasets commonly used in the field of semi-supervised semantic segmentation.

Category	Datesets	Classes	Resolution	Annotated Samples	Domain Coverage	Download
Natural Images	Pascal VOC [188]	21	500×375	>16,000	General objects	Link
	ADE20K [189]	150	Varying	>25,000	Multiple (indoor, outdoor)	Link
	MS-COCO [190]	80	Varying	>200,000	Multiple (everyday scenes)	Link
Street-view Images	KITTI [191]	28	1242×375	400	Autonomous driving	Link
	CamVid [192]	32	960×720	701	Road scenes	Link
	Cityscapes [17]	19	2048×1024	5,000	Urban Driving Scenes	Link
	RainCityscapes [193]	32	2048×1024	5,000	Rainy weather street scenes	Link
	FoggyCityscapes [194]	32	2048×1024	5,000	Foggy street scenes	Link
	BDD100K [195]	40	1280×720	10,000	Diverse road scenes	Link
	Mapillary [196]	66	Varying	25,000	Road scenes	Link
Medical Images	MoNuSeg [197]	2	Varying	30	Nuclei segmentation	Link
	Pancreas-NIH [151]	2	512×512	82	Pancreas (CT)	Link
	LA [159]	2	Varying	100	Left atrium (MRI)	Link
	ACDC [152]	4	256×256	100	Cardiac MRI	Link
	Kvasir-SEG [170]	2	Varying	1,000	Gastrointestinal (polyp detection)	Link
	BRATS [156]	5	Varying	2,040	Brain tumor (MRI)	Link
	ISIC-2018 [165]	7	Varying	3,694	Skin lesion (dermatology)	Link
	NCT-CRC-HE-100K [167]	9	224×224	100,000	Colorectal cancer histology	Link
Satellite Images	GID [63]	5	7200×6800	150	Land cover classification (satellite)	Link
	iSAID [176]	15	800×800	2,806	Aerial imagery	Link
	FloodNet [198]	9	4000×3000	3,200	Flood land related	Link
	xBD [178]	4	1024×1024	22,068	Building damage detection	Link
	EuroSAT [199]	10	64×64	27,000	Diverse ground cover	Link

labeled samples are scarce. In their latest study, [175] proposed an innovative method that leverages bicommutative entropy consistency and a teacher-student framework. The complexity of this task lies in the presence of multiple categories, intricate topography, significant category overlap, and ambiguous features. Therefore, the authors integrated the channel attention (CA) mechanism into the teacher coding network, which effectively filters the feature mapping and mitigates the noise interference, thus refining feature extraction and reducing the information entropy generated by the coding network.

IV. Classical Datasets

This section describes the datasets suitable for semantic segmentation tasks and categorizes them according to their content and image characteristics.

Public datasets are an important resource for the research community as they shape and encapsulate the task's challenges and serve as benchmarks for evaluation. However, large-scale semantic segmentation datasets are still relatively scarce. By 'large-scale' we mean datasets with high capacity (100,000 sheets or more) and high resolution (1024x1024 pixels or higher). On the one hand, many existing datasets, while valuable, tend to be domain-specific, focus on a limited number of scenes, and lack diversity because imaging equipment varies from scene to scene. On the other hand, the cost of annotating pixel-level data for semantic segmentation is prohibitive compared to other computer vision tasks (e.g., image classification [202]–[205], which requires only image-level annotations, or object detection [206]–[208], which requires labels the category and location of each object). This high cost further limits the availability of large-scale datasets with rich and fine-grained annotations.

Table III lists the most commonly used semi-supervised datasets for different tasks. It is worth noting that these datasets listed in the table can also be used as benchmarks in fully supervised, weakly supervised, or unsupervised situations. This is simply a matter of selectively using the data and corresponding labels during training. Since this study focuses on SSSS, some fully labeled images are usually selected for SSL, in the proportion of $5\%, 10\%$, etc., and the rest are left unlabelled. From Table III we can see that there are a limited number of datasets that meet the large-scale and high-resolution criteria. Therefore, we hope that future research should prioritize the development of more comprehensive datasets to support the training of models in more diverse and complex scenarios, including extreme lighting conditions and finer object annotations. These larger datasets will improve model performance, especially when pre-training for real-world applications.

V. Experiment

In this section, we provide a qualitative comparison and a quantitative assessment of representative conventional methods covered in our survey.

A. Experimental Setup

Dataset. When choosing the experimental datasets for evaluation, we took into account the range of categories and distinctive features. This approach enabled a more effective quantitative analysis and qualitative discussion while also allowing for a comparison of the different methods described in the prior section. Initially, we select the PASCAL VOC 2012 dataset [188], [209], as it is the most frequently utilized in the SSSS domain, to perform experiments aimed at quantitatively comparing different approaches. This dataset encompasses a wide variety of natural images and categories, which aids in achieving more reliable experimental outcomes. To ensure consistent evaluation of the algorithm's performance, for the

TABLE IV: Detailed information about the experimental data. Includes specific divisions as well as labeling ratios and the number of labeled and unlabeled images in the constructed partitions.

Datasets	Classes	Full Train Set	Test Set	Val Set	Label Ratio	Labeled Images	Unlabeled Images
PASCAL VOC 2012 [188]	21	10582	4952	1449	1 / 2	5291	5291
					1 / 4	2646	7936
					1 / 8	1323	9259
					1 / 16	662	9920
Cityscapes [17]	19	2975	1525	500	1 / 4	744	2231

PASCAL VOC 2012 dataset, we set the crop size to 321 × 321 and trained for 80 epochs. We employ a learning policy starting with an initial learning rate of 0.001, which is then adjusted by multiplying it with $\left(1-\frac{epoch}{total\_epoch}\right)^{power}$. The power and weight decay are set to 0.9 and 0.0001, respectively. Secondly, because the images have high resolution and contain many categories within each image (unlike PASCAL VOC 2012, which emphasizes fewer categories per image), we opted for the Cityscapes dataset for visualizing qualitative experiments. This approach allows us to assess the ability of trained models in complex environments with multiple adjacent categories that may be similar or share semantic links.

Datasets Partition. In SSL experiments, dataset partition is a key aspect that affects performance, and to obtain a comparable and more plausible result with others, our quantitative experiments on PASCAL VOC 2012 used the partitioning protocol of U2PL [64], a commonly used data partitioning protocol that encompasses a wide variety of scenarios in terms of label ratios. The whole training was divided into two subsets, where 1/2, 1/4, 1/8, and 1/16 scale data were selected as labeled sets and the rest as unlabeled sets. In contrast, our qualitative experiments on the Cityscapes dataset uniformly chose the 1/4 labeled set for training evaluation. The validation methods used in the experiments involve the typical approach of splitting each dataset into training and validation subsets. Our standard validation approach includes a simple holdout, a training set, and a validation set. Detailed descriptions of partitions can be found in Table IV.

Backbone Network and Specific Settings. Different semi-supervised methods rely on various underlying models and backbone networks. Consequently, the final performance of these segmentation methods is highly dependent on the network, complicating performance comparisons. To address this, we standardized this critical aspect in our experiments. Based on setups and results from the literature, we selected DeepLabV3+ [201] as the base model and ResNet101 [210] as the backbone, a combination that yields superior performance. All experiments in this review were conducted using Python as the programming language and the PyTorch deep learning framework. Specific quantitative experiments were trained and tested on 4 NVIDIA V100 GPUs with a batch size of 16.

Performance Metric. The performance metric we use in this experiment is the standard evaluation metric in SSL, which is the mean intersection over union ($mIoU$). Unlike accuracy metrics commonly used for classification tasks, $mIoU$ can be robust to the presence of unbalanced classes, which is very common in problems with pixel-level labeling. Specifically, $mIoU$ measures the ratio of the number of true positives (TP) to the sum of true positives (TP), false positives (FP), and false negatives (FN), averaged as shown in the formula below:

$$mIoU = \frac{1}{N}\sum_{i=1}^{N}\frac{N_{ii}}{\sum_{j=1}^{N}N_{ij}+\sum_{j=1}^{N}N_{ji}-N_{ii}} \tag{8}$$

where N is the number of categories, N_{ii} is the number of true positives (TP) for category i, N_{ij} is the number of false positives (FP) for categories i and j, and N_{ji} is the number of false negatives (FN) for categories j and i.

Method Selection. The purpose of the selection phase in our experimental methodology was to take into account all viewpoints and choose 1-4 studies in each subcategory for assessment using the same parameters and protocols (quantitative evaluation), as well as to engage in discussion based on the experimental outcomes and related approaches from the original literature (qualitative discussion). As our baseline, we utilized the original self-training model [33], which was trained solely with labeled data. From the model perspective, GIST&RIST [86], USCS [32] and ST++ [33] were selected for single-model, while DMT [96], CPS [89] and CCVC [100] were chosen for multi-model approaches. Regarding the pseudo-label refinement perspective, CARD [79], CISC-R [88], LogicDiag [101], and Corrmatch [71] focus on "Label Update". On the other hand, CAFS [97], TorchSemiSeg2 [105], PGCL [106], and GTA-Seg [107] are discussed around "Filter-only". From the data augmentation perspective, we chose three methods for traditional augmentation, CutMix [108], AugSeg [112], and iMAS [113], the latter two mainly exploring randomized combinations and instance-specific augmentation on a traditional basis. As for the optimization methods, PS-MT [133], Corrmatch [71] focus on the loss function perspective, while PGCL [106], ESL [69] and CPCL [118] improve the training from the training strategy perspective. In addition, we assess the USRN [180], AEL [179], CFCG [138], and UniMatch [124] as part of hybrid approaches that integrate CR. Additionally, DGCL [139] and U2PL [64] are merged with CL, and ADS-SemiSeg [115] is merged with AT. It is important to acknowledge that the methodologies within each category are designed to address specific aspects of PL in SSSS and to overcome particular challenges. Model perspective primarily assesses methods in terms of their framework structure. Pseudo-label refinement primarily enhances the quality of pseudo-labels. Data augmentation improves the model's robustness by increasing data diversity. Optimization improvement stabilizes the training process through consistency and loss function. Therefore,

TABLE V: Results of various PL methods on the PASCAL VOC 2012 val dataset and self-training baseline results. Each column corresponds to the ratio of labeled/unlabeled images (the numbers in parentheses indicate the number of labeled images used in each scenario). In each partition, the results obtained with the best method, respectively, are bolded. (Indicator: $mIoU$).

Category		Method	1 / 2 (5291)	1 / 4 (2646)	1 / 8 (1323)	1 / 16 (662)
Baseline			**75.13**	**72.15**	**69.67**	**64.40**
Model Structure	**Single Model**	GIST&RIST*	—	—	70.76	—
		ST++˜	76.53	75.51	74.97	72.43
		USCS*	78.63	77.09	76.20	74.52
	Mutual Model	DMT*	—	71.80	71.00	—
		CPS	78.64	77.68	76.44	74.48
		CCVC	—	79.00	78.40	77.20
Pseudo-label Refinement	**Label Update**	CARD*	—	—	74.07	—
		CISC-R˜	78.02	77.36	77.25	75.39
		LogicDiag	81.00	**80.62**	80.24	**79.65**
		Corrmatch˜	78.73	78.86	78.28	76.87
	Filter-only	CAFS*	80.70	79.20	77.70	75.10
		TorchSemiSeg2	77.04	76.80	74.91	73.38
		PGCL*	—	76.80	76.80	73.60
		GTA-Seg	**81.01**	80.57	80.47	77.82
Data Enhancement	**Classic Data Enhancement**	CutMix-Seg˜	75.89	74.25	72.69	72.56
		AugSeg	—	80.5	**81.46**	79.29
		iMAS	—	79.30	78.40	77.20
	Geometric Data Enhancement	ADS-SemiSeg	—	74.90	73.40	—
Optimization Improvement	**Loss Function**	PS-MT	79.76	78.72	78.20	75.50
		Corrmatch˜	78.73	78.86	78.28	76.87
	Training Strategy	PGCL*	—	77.90	76.80	73.60
		ESL*	79.98	79.02	78.57	76.36
		CPCL	75.30	74.58	73.74	71.66
Hybrid Techniques	**Combining CR**	USRN*	—	—	—	72.30
		AEL	80.29	78.06	77.57	77.20
		CFCG*	80.77	**80.42**	**79.40**	**77.39**
		UniMatch	—	77.20	77.10	76.50
	Combining CL	DGCL*	**80.96**	79.31	78.37	76.61
		U2PL˜	79.94	78.70	77.60	74.43
	Combining AT	ADS-SemiSeg	—	74.90	73.40	—

[1] ˜ indicates that we reproduced the results.
[2] * means that the article does not disclose the open source code, and we cite the results given in the paper, which are only used here for comparison, where the GIST & RIST use the DeepLabV2 [87] segmentation model and the results of DMT are replicated in [57].

the methods within different categories are not in direct competition; rather, they are complementary strategies for addressing distinct challenges. For instance, although AugSeg demonstrates efficacy in the data enhancement category, this does not imply that it is superior to CAFS in the pseudo-label optimization category.

B. Results and Discussions

In this part, we present and analyze the results we have obtained. Initially, we display and evaluate the quantitative outcomes derived from the chosen traditional methods on the PASCAL VOC 2012 and perform an analysis for assessment. Secondly, we present the outcomes achieved in urban environments, providing a qualitative and visual examination of several widely used techniques on the CityScapes.

1) Quantitative results on PASCAL VOC 2012:

We present in Table V the results of qualitative experimental evaluations of different methods under the same conditions on the PASCAL VOC 2012 dataset, encompassing all categories segmented in our study. Our baseline is the most straightforward single-model self-training PL. This process includes training a model on labeled data, utilizing it to generate predictions for unlabeled data to create pseudo-labels, and subsequently retraining the model with the complete dataset. From a broad perspective, partitioning affects the performance of all methods, including the baseline. As the amount of labeled data decreases, performance is affected and subsequently declines. The initial evaluation perspective focuses on the difference between each method and the baseline. The improvement of each method over the PL is justified.

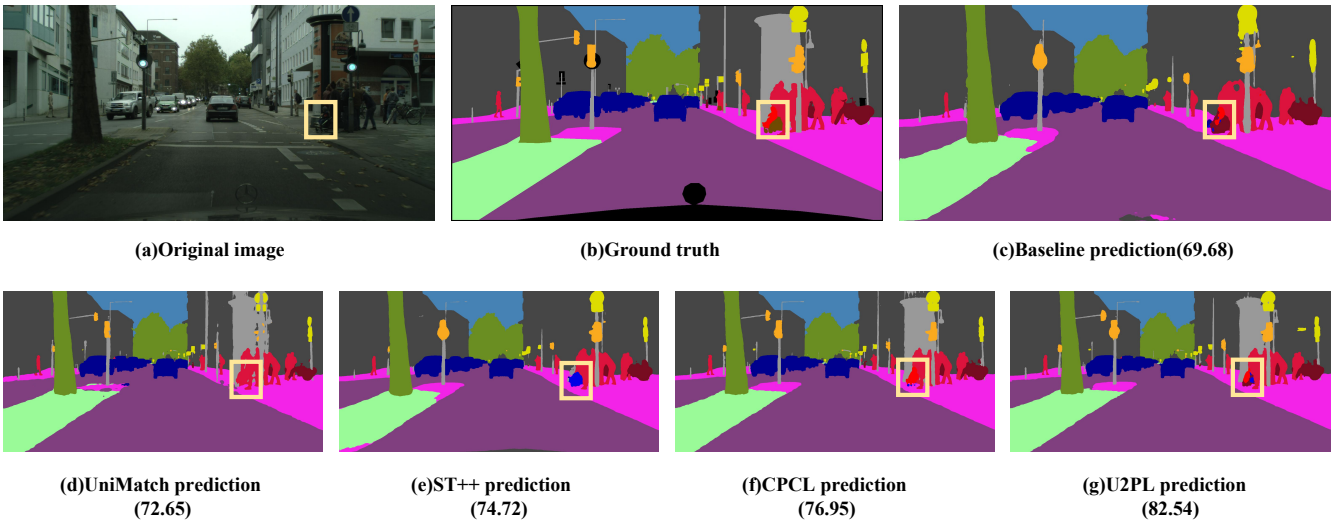
Model Structure. As shown in Table V, the "Model Structure" category indicates that "Mutual Model" structures (e.g., CPS and CCVC) mostly outperform the single model as the labeled/unlabeled ratio decreases. For example, when the ratio is 1/16, CCVC has a $mIoU$ of 77.20, which outperforms the single model by 2-5%. This suggests that the mutual learning framework can benefit from the integration of complementary information between models, especially when labeled data is limited. We speculate that the ability of mutual models to refine pseudo-labels and adjust predictions across multiple networks may help reduce errors caused by noisy pseudo-labels. Notably, individual models still perform well when using high-quality labeled data. For example, USCS achieves

78.63 $mIoU$ at 1/2, indicating that the single model can still achieve competitive results with sufficient labeled data. However, in low-labeling environments, single models may face challenges due to the lack of collaborative information refinement, whereas mutual aid models offer significant performance advantages by exploiting inter-model consistency, and their ability to handle noisy pseudo-labels makes them particularly suitable for extreme SSL scenarios. However, the computational burden associated with running multiple models should be fully considered when deploying them in practice.

Pseudo-label Refinement. Refinement techniques such as LogicDiag and CAFS have proven to be effective in dealing with noisy labeling challenges and improving the quality of pseudo-labels with varying labeled/unlabeled ratios. However, label-update still outperforms filter-only on average. In particular, LogicDiag achieves an optimal performance of 80.62 $mIoU$ at 1/4, and even at more challenging settings (e.g., 1/16), LogicDiag shows a strong performance of 79.65 $mIoU$, underscoring its effectiveness in correcting erroneous pseudo-labels. It is worth noting that the pure filtering approach also performs well in some scenarios, but when the labeling is very sparse, the performance is adversely affected. This suggests that in the case of highly unlabeled data, filtering strategies may not be sufficient to address the noise problem. Refinement of the techniques is essential to maintain a high quality of pseudo-labels and to ensure consistency of performance across categories. In addition, combining these techniques with other filtering strategies can increase their robustness.

Data Enhancement. The results show that data enhancement strategies can significantly improve performance. Of the traditional augmentation methods, AugSeg shows the most promise for improving model performance under a variety of labeled/unlabeled ratios. It provides a more robust method for increasing data diversity by applying random selection augmentation than simpler augmentation strategies such as CutMix-Seg. When labeled data becomes scarce, at 1/16, AugSeg continues to prove its superiority by maintaining a higher $mIoU$ value (79.29 $mIoU$) than other classical methods. Geometric data enrichment focuses specifically on applying geometric transformations to enrich the spatial configurations encountered by the model during training. In terms of geometric enhancement, the ADS-SemiSeg method achieves a $mIoU$ of 73.40 at 1/8, which is a 3.73% improvement over the Baseline, an improvement that suggests that exploring the spatial diversity of an image can lead to some performance gains, a factor often overlooked by traditional methods. These results suggest that while it is still valuable to investigate traditional enhancement, there is no doubt that geometric transformations can help further expand the diversity of training examples for more effective generalization, which is currently underexplored, and future work should delve deeper into exploration and geometric enhancement, as well as hybrid methods combining traditional and geometric enhancement.

Optimization Improvement. The results of PS-MT and Corrmatch demonstrate the efficacy of specialized loss functions and consistency-based training strategies. PS-MT achieves 79.76 $mIoU$ at 1/2, illustrating the value of employing the mean-teacher framework to regularize the noisy predictions. Corrmatch exhibits robust performance at diverse ratios, particularly at the 1/16 ratio, attaining 76.87 $mIoU$ and outperforming numerous competing methods. The training strategy also plays a significant role in enhancing the robustness of the model. ESL attains a $mIoU$ of 79.02 at the 1/4 ratio, a value that is only 1.6 units away from the optimal result. This outcome suggests that a course-learning strategy that gradually increases training difficulty is highly effective in semi-supervised scenarios. Optimization improvements, especially through tailored loss functions and strategies, ensure model robustness. These methods are essential for reducing overfitting to noisy pseudo-labels and thus improving generalization to unknown data. They should therefore be considered an important part of the semi-supervised training



(a)Original image **(b)Ground truth** **(c)Baseline prediction(69.68)**

(d)UniMatch prediction (72.65) **(e)ST++ prediction (74.72)** **(f)CPCL prediction (76.95)** **(g)U2PL prediction (82.54)**

Fig. 6: Visual evaluation of example images in the CityScapes dataset using the UniMatch, ST++, CPCL, and U2PL methods with pedestrian, vehicle, road, sidewalk, vegetation, and building categories as the main representatives. Parentheses after the methods indicate example IoU results, and we also show raw image, ground truth, and the Baseline results.

pipeline.

Hybrid Techniques. It is evident that hybrid techniques that integrate PL with CR, CL, or AT demonstrate considerable potential. USRN and AEL integrate CR to solve the class imbalance problem. AEL achieves 80.29 $mIoU$ at the 1/2 ratio, which serves to highlight its ability to balance the training of performing and underperforming classes adaptively. Furthermore, hybrid techniques such as DGCL and U2PL integrate CL to yield notable feature representation enhancements, as DGCL attains 80.96 $mIoU$ at 1/2. These approaches not only enhance the model's capacity to process noisy pseudo-labels but also facilitate the model's ability to learn meaningful feature representations, even when labeled data is scarce. Hybrid techniques that integrate multiple strategies are highly effective in improving model performance in SSL environments. They provide a complementary approach, as evidenced by the effectiveness of techniques such as AEL and DGCL in complex real-world applications with limited labeled data.

The results presented in Table V indicate that no single technique can be considered a universal solution to the challenges faced by SSSS. Nevertheless, the use of hybrid techniques, which combine multiple approaches, has consistently demonstrated superior performance compared to the use of single techniques. These methods enhance the efficacy of PL by optimizing feature representation, addressing class imbalance, and mitigating the impact of unlabeled data. In particular, pseudo-labeling refinement techniques like LogicDiag and Cormatch, when integrated with advanced enhancement and hybrid training strategies, represent a promising avenue for future research in PL techniques. These methods offer robust performance improvements, particularly in scenarios with high ratios of unlabeled to labeled data, and are well suited for practical applications in real-world segmentation tasks.

2) Qualitative results on Cityscapes:

In this subsection, we utilize the CityScapes dataset for qualitative assessment and visualization to better compare the segmentation outcomes of several state-of-the-art methods. The methods evaluated include UniMatch, ST++, CPCL, and U2PL. The detailed visualization results are presented in Fig. 6 and Fig. 7. Additionally, we compare these results with ground truth and baseline methods. Subsequently, we will conduct a detailed analysis and evaluation of the results obtained from the different segmentation methods.

In the initial visualization example depicted in Fig. 6, it is evident that some of the predominant categories such as pedestrians, vehicles, vegetation, and buildings are fairly well represented. However, infrequent categories like strollers exhibit more noticeable misclassifications and confusion. Additionally, the Mailbox (gray) category is often confused with the Utility Pole category in UniMatch and is not correctly identified by the baseline and ST++ methods. For classes that are small but numerous, such as poles and streetlights, although all methods can detect the presence of poles, they are not consistently accurate in pinpointing each pole's location. This can be attributed to the fact that models find it easier to segment classes occupying larger areas in the image, whereas classes occupying smaller and more fragmented areas pose greater prediction challenges and are more likely to be confused with adjacent classes, leading to lower IoU and inaccurate class prediction results.

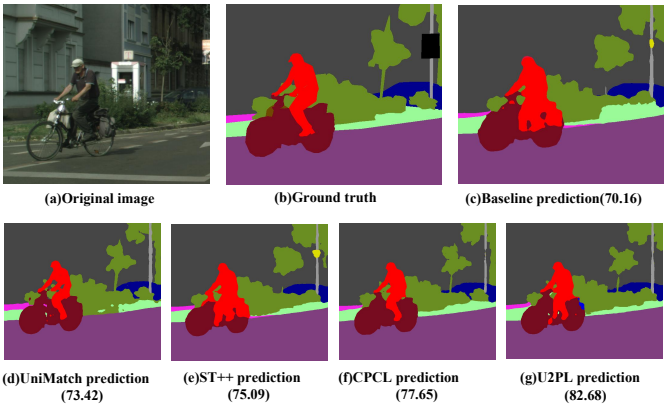


Fig. 7: Visualization evaluation using UniMatch, ST++, CPCL, and U2PL methods on example images in the Cityscapes dataset.

In the second visualization example illustrated in Fig. 7, it is evident that the predominant classes in the image are riders and bicycles, while the minor courses are poles and traffic signs. The methods exhibit similarly excellent performance in predicting vegetation, vehicles, roads, and buildings; however, the prediction performance for riders is generally subpar. Although there is no strict distinction between the rider and human categories, the difference arises because riders are typically adjacent to bicycles or motorcycles. Consequently, the model must discern the semantic contextual information between these categories. In other words, the model needs to understand the intra-class and inter-class semantic relationships of pixels predicted to belong to the rider category; otherwise, it may confuse this category with the person category.

3) Distribution by Category on PASCAL VOC 2012:

In recent studies [33], [95], images from a high-quality original training set were extracted and labeled. The experiments presented in this subsection are based on this case, where 1,464 images were extracted as labeled images, and the remainder were labeled as unlabeled. As illustrated in Fig. 8, we present the visualization outcomes of conventional supervised methodologies and classical PL techniques (Cormatch [71] as a case in point) on the PASCAL VOC with disparate class data distributions, which IoU assesses as a The evaluation metric demonstrates that the PL methods align with the supervised learning methods in terms of the overall trend, as illustrated in the figure. Notably, the PL exhibits superior performance in several classes, particularly in an imbalanced data distribution between classes. To illustrate, the IoU for the airplane and cow classes demonstrates a notable enhancement under the PL. Conversely, certain courses with high degrees of similarity (e.g., cat and dog) exhibit a narrower gap between the supervised and PL, which can be attributed to the inherent complexity of the training data. Furthermore, the figure's fluctuations demonstrate the sensitivity of different categories in data distribution. In the case of sparser category data, the PL method can effectively utilize unlabelled data by

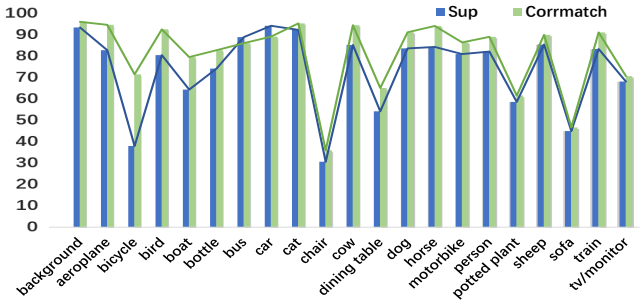


Fig. 8: The illustration provides a comprehensive representation of the trends in category data distribution for the experimental results of the supervised and classical PL methods (in the case of Corrmatch [71]) under the PASCAL VOC 2012 val dataset. (Indicator: IoU). Labeled images are selected from **original** training set. (best viewed in color).

generating pseudo-labels, thereby compensating for the effect of insufficient labels in supervised methods.

VI. Challenges and Opportunities

This section highlights some of the most promising and valuable PL research avenues in the domain of SSSS, as shown in detail in Fig. 9.

A. Quality enhancement with foundation models

Although current PL methods have made significant progress in typical SSSS, the choice of their underlying models is still limited, often relying on generic architectures such as ResNet [210] or Vision Transformer [211]. These architectures, although highly capable of feature extraction, are not optimized for the specific challenges of the pseudo-label generation process. For example, existing methods are usually based on static feature extraction and label generation, lacking dynamic tuning and interactive feedback mechanisms, whereas the accuracy of pseudo-label generation is critical to the overall performance of semantic segmentation. In the initial training phase, incorrect pseudo-labels may cause the training model to fall into local optimum and reduce overall performance.

To address these issues, new techniques such as interactive cues in combination with base models such as Segment Anything Model (SAM) [212] are used. In existing research, SAM can generate high-quality seeds or cues applied to weakly- [213] or un-supervised [214] segmentation. At the same time, SAM fine-tunes the behavior of the model through interactive cues, which can provide more flexibility and semantic information support for pseudo labels generation, and has now been applied as a guide to the task of SSL referring expression segmentation [215]. While the application level has only been used as an enhancement aid module for medical image segmentation [216]. Future research on PL in SSSS can explore how to combine SAM as an auxiliary model with the iterative pseudo-label generation process and use its dynamic cueing function to further improve the performance of online SSSS by gradually refining from the initial rough pseudo-label combined with the recurrent strategy.

B. Multimodal Fusion

So far, most existing PL in SSSS techniques often face problems such as inter-class imbalance [180], boundary ambiguity [217], and pseudo labels noise [59]. These challenges significantly affect the classification accuracy of the models in complex scenarios. In addition, the process of PL relies on high-confidence regions, leading to the under-utilization of unlabelled data. Although a teacher model [84] is introduced to improve the quality of pseudo labels, it is still challenging to scale this approach across different data modalities. Future research can draw on the development of cross-modal learning and multimodal fusion techniques [218]–[221] to improve pseudo labels quality and segmentation performance by introducing multimodal data (depth maps, infrared images, text descriptions). For example, significant progress has been made in multimodal fusion in target detection [222] and image-text alignment tasks [223]. Different modalities can provide complementary semantic information, effectively enhancing the robustness of the model when a single modality is limited.

Designing modular networks with a unified encoder structure [224] can also be a solution because it has shown its advantages for fine-grained semantic processing in other vision tasks as well. For example, in text-video segmentation [225] and classification tasks [226], modular networks can better capture local information and perform overall optimization by processing different modalities or sub-tasks separately, without the need to design separate models for each modality. Introducing it into pseudo-label generation can improve problems such as boundary-blurring by adaptive processing of different regions or semantic categories, and the model can use complementary multimodal information to generate more accurate pseudo-labels. Especially when segmenting complex scenes, the multimodal fusion strategy [227] is expected to adjust the segmentation accuracy for specific regions based on cues or interactions, enhance the collaborative reasoning ability between different modalities, and further improve the overall quality of pseudo-labels. In addition, inspired by the Large Language Model (LLM) [228], new directions of semantic alignment and fusion can be explored in the future to optimize the pseudo labels generation process further.

C. Domain Adaptation

In laboratory environments, current PL methods mainly rely on the assumption that the training and test data are in the same distribution (same domain), which has the advantage of ensuring performance consistency when the trained model is applied to the test. In practice, however, different environments (e.g., weather changes, lighting conditions, fog conditions, camera angles, etc.) can lead to significant domain differences, which usually trigger a degradation in the quality of the pseudo-labels and thus affect the performance of the model. Therefore, learning how to address the large domain gaps between data from different sources is one of the key issues to improve the generalisability of models.

Some existing works have attempted to narrow the gap between different data distributions through domain adaptation [229] techniques. For example, [230] has been conducted to

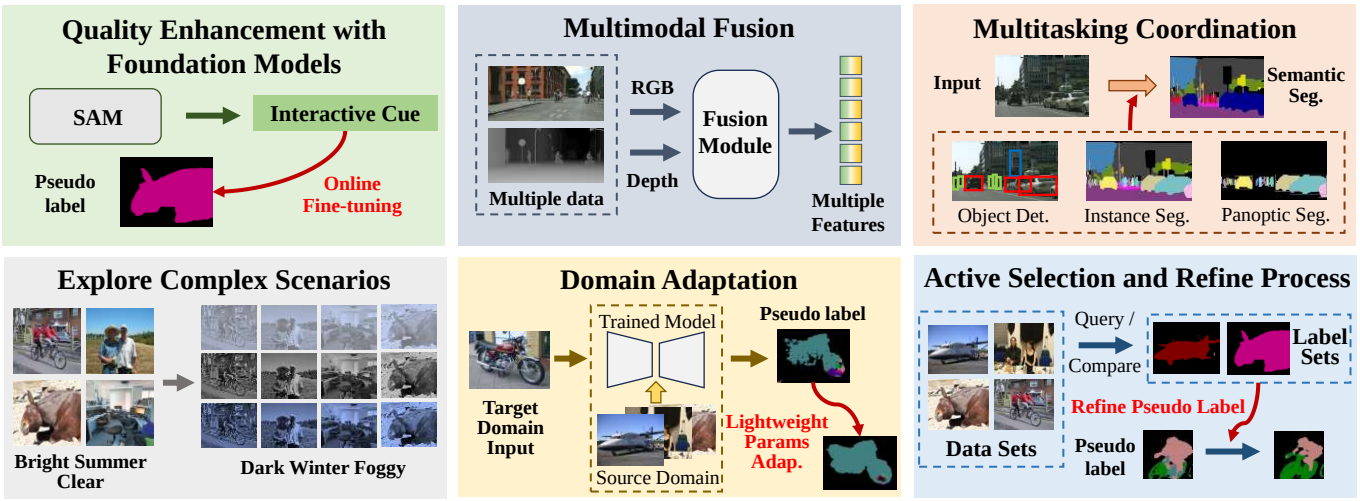


Fig. 9: This illustration details six possible future research directions for PL technology in SSSS for its future development and practical application (best viewed in color).

reduce the distributional differences between different data domains through adversarial learning and self-supervised. However, most approaches rely on a large amount of computational resources and their effectiveness is still limited when dealing with highly variable domains. Recently, pre-trained visual models have shown great potential in domain migration tasks [231]–[233]. These models have stronger generalization capabilities due to training on large-scale, diverse data and can help cope with the domain drift problem. Some approaches apply pre-trained models to specific tasks through full fine-tuning techniques [234], but this triggers high computational overheads and can compromise the generalization capabilities contained in pre-trained models. In the future, a domain adaptive approach based on lightweight prompt learning strategies may be an effective research path [235]. This approach adapts to new domains at a very small computational cost by introducing a small number of learnable parameters, while retaining the knowledge of the pre-trained model, and has been preliminarily validated in tasks such as object detection [236] and cross-modal learning [237]. In PL of SSSS, guiding the pseudo-label generation process through the cue learning strategy can better cope with the domain migration problem, and improve the performance in the new domain.

D. Explore Complex Scenarios

Most PL techniques efforts focus on relatively simple, existing standard datasets, such as PASCAL VOC [209], which are usually of moderate size and have a single scene. In practice, it is often necessary to deal with more complex scenes that include not only diverse objects and backgrounds, but may also be accompanied by more occlusions, extreme lighting variations, and dense object interactions, which place higher demands on existing PL methods. So researchers have begun to explore more challenging datasets and scenarios. For example, large-scale datasets such as Mapillary Vistas [196] and ADE20K [189] provide test benchmarks for complex scene segmentation. The team building Cityscapes [17] has similarly introduced the datasets Raincityscapes [193] in rain and Foggycityscapes [194] in fog, which, compared to common These datasets contain more diverse scenes, more types of objects, and higher resolution than common, and better simulate the actual environments in the real world. Although the application of PL on these datasets is still limited, some preliminary studies have shown their potential. [132] achieved preliminary results by generating pseudo-labels and performing adaptive learning on large-scale datasets. However, due to the low accuracy of pseudo-labels on large-scale complex datasets, exploring SSSS in more diverse scenarios [238] still has great research potential.

E. Multitasking Coordination

Segmentation in practice often needs to work in concert with other visual tasks (detection, classification) to achieve a more comprehensive visual understanding. However, in the pseudo-label generation, most of the current generated labels are usually optimized for a single task only (e.g., SSSS), neglecting information sharing and co-optimization with other tasks. Given that multi-task learning [239] has shown its potential in other visual tasks. [240] significantly improved the performance of both detection and segmentation tasks by jointly learning them through multitasking.

Considering that in complex scenes, single-task pseudo-label generation may be inaccurate due to the lack of contextual information, it makes sense to explore how to provide more accurate and comprehensive semantic information through a multi-task learning framework. Some recent works have made valuable attempts. For example, e.g., co-training semantic segmentation with tasks such as target detection and scene understanding [201] enables the model to generate more accurate pseudo-labels based on shared feature representations and using output information from other tasks. In the future, PL can draw on the strategy of multi-task coordination [241], which is expected to generate more accurate pseudo-labels and improve the accuracy and robustness of segmentation through multi-tasks sharing valuable and complementary contextual information. Multi-task optimization frameworks proposed in

recent years, such as Quadronet [242] and UPerNet [243], have shown their potential in visual tasks. These frameworks enable information transfer between tasks by sharing convolutional features or Transformer modules.

F. Engage in the active selection and refine the process

Although some typical approaches have achieved encouraging results, there are still some limitations and challenges in the current PL generation process. Since the generation of pseudo labels is based on model prediction, if the initial performance of the model is poor, the quality of the corresponding pseudo-labels will also be affected [59]. In addition, pseudo-labels generated in hard-to-distinguish regions or class-imbalanced scenarios often contain errors [180], which may negatively affect the model training.

To better address this problem, Active Learning (AL) [244] may be a very relevant solution to the SSL task, which, instead of training the model on the entire dataset, focuses on selecting a subset of the most informative data points for requesting additional labels, allowing the model to learn from the most valuable examples efficiently and cost-effectively. AL has been applied to tasks such as image classification [245], object detection [246], etc. In the future, AL-based active selection and pseudo labels refinement strategies can be combined to pool resources more effectively, which in turn improves the quality of pseudo-labels and avoids the interference of incorrect labels on model training.

VII. Summary

This review systematically summarizes and categorizes pseudo-labeling solutions in the domain of semi-supervised semantic segmentation over recent years. It integrates various methods and their enhancements and also provides an overview of hybrid techniques combining pseudo-labeling with other semi-supervised segmentation strategies. The paper draws reliable conclusions from both quantitative experiments and qualitative assessments, highlighting the remaining challenges and potential future directions in this field. Additionally, it briefly discusses some emerging approaches in the current landscape of semi-supervised segmentation.

After thorough analysis, we conclude that enhancing the current PL methods in SSSS can be approached from the model perspective, the label refine perspective, the data perspective, and optimization perspective. Our experiments indicate that generating various hybrid method combinations holds great promise, particularly when combining consistency regularisation with contrast learning. Methods for refining pseudo-labels have also demonstrated outstanding results. Furthermore, enhancing data perturbation and optimizing training strategies are promising avenues for future advancement.

Although current PL techniques have advanced significantly in SSSS, their effectiveness remains constrained in complex scenarios, particularly when dealing with domain shifts or extreme conditions. In such cases, the generalization capacity and precision of pseudo-labels are often lacking. We highlight the importance of investigating complex scenarios and discuss strategies to enhance pseudo-label quality via active learning, domain adaptation, and multimodal data fusion. Given the challenges of handling complex real-world situations with current pseudo-labeling and optimization techniques, future PL approaches need to develop more adaptable and intelligent methods to actively select the most representative data samples and adaptively adjust them by integrating inter-domain variations to boost model robustness and generalization. Additionally, considering training costs, we propose that exploring more efficient network architectures and multitasking approaches is valuable for future applications in specific domains.

References

[1] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image segmentation using deep learning: A survey," *TPAMI*, vol. 44, no. 7, pp. 3523–3542, 2021.

[2] H. Huang, S. Xie, L. Lin, R. Tong, Y.-W. Chen, Y. Li, H. Wang, Y. Huang, and Y. Zheng, "Semicvt: Semi-supervised convolutional vision transformer for semantic segmentation," in *CVPR*, 2023, pp. 11 340–11 349.

[3] L. Ran, Y. Zhang, and G. Hua, "Cannet: Context aware nonlocal convolutional networks for semantic image segmentation," in *ICIP*, 2015, pp. 4669–4673.

[4] R. M. S. Bashir, T. Qaiser, S. E. A. Raza, and N. M. Rajpoot, "Consistency regularisation in varying contexts and feature perturbations for semi-supervised semantic segmentation of histology images," *Medical Image Analysis*, vol. 91, p. 102997, 2024.

[5] H. Li, D.-H. Zhai, and Y. Xia, "Erdunet: An efficient residual double-coding unet for medical image segmentation," *TCSVT*, 2023.

[6] Z. Xu, Y. Wang, D. Lu, L. Yu, J. Yan, J. Luo, K. Ma, Y. Zheng, and R. K.-y. Tong, "All-around real label supervision: Cyclic prototype consistency learning for semi-supervised medical image segmentation," *JBHI*, vol. 26, no. 7, pp. 3174–3184, 2022.

[7] X. Zhang, Z. Zhao, L. Ran, Y. Xing, W. Wang, Z. Lan, H. Yin, H. He, Q. Liu, B. Zhang, and Y. Zhang, "Fasticenet: A real-time and accurate semantic segmentation model for aerial remote sensing river ice image," *Signal Processing*, vol. 212, p. 109150, Nov. 2023.

[8] L. Ran, L. Wang, T. Zhuo, Y. Xing, and Y. Zhang, "Ddf: A novel dual-domain image fusion strategy for remote sensing image semantic segmentation with unsupervised domain adaptation," *TGRS*, 2024.

[9] J. Geng, S. Song, and W. Jiang, "Dual-path feature aware network for remote sensing image semantic segmentation," *TCSVT*, 2023.

[10] W. Wang, L. Ran, H. Yin, M. Sun, X. Zhang, and Y. Zhang, "Hierarchical shared architecture search for real-time semantic segmentation of remote sensing images," *TGRS*, vol. 62, pp. 1–13, 2024.

[11] W. Liu, W. Li, J. Zhu, M. Cui, X. Xie, and L. Zhang, "Improving nighttime driving-scene segmentation via dual image-adaptive learnable filters," *TCSVT*, vol. 33, no. 10, pp. 5855–5867, 2023.

[12] L. Wang, D. Li, H. Liu, J. Peng, L. Tian, and Y. Shan, "Cross-dataset collaborative learning for semantic segmentation in autonomous driving," in *AAAI*, vol. 36, no. 3, 2022, pp. 2487–2494.

[13] J. Li, H. Dai, H. Han, and Y. Ding, "Mseg3d: Multi-modal 3d semantic segmentation for autonomous driving," in *CVPR*, 2023, pp. 21 694–21 704.

[14] M. Cheng, L. Hui, J. Xie, and J. Yang, "Sspc-net: Semi-supervised semantic 3d point cloud segmentation network," in *AAAI*, vol. 35, no. 2, 2021, pp. 1140–1147.

[15] L. Zhao and W. Tao, "Jsnet++: Dynamic filters and pointwise correlation for 3d point cloud instance and semantic segmentation," *TCSVT*, vol. 33, no. 4, pp. 1854–1867, 2022.

[16] J. Liu, Y. Chen, B. Ni, and Z. Yu, "Joint global and dynamic pseudo labeling for semi-supervised point cloud sequence segmentation," *TCSVT*, 2023.

[17] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *CVPR*, 2016, pp. 3213–3223.

[18] A. Vezhnevets, V. Ferrari, and J. M. Buhmann, "Weakly supervised structured output learning for semantic segmentation," in *CVPR*, 2012, pp. 845–852.

[19] J. Chen, W. Lu, Y. Li, L. Shen, and J. Duan, "Adversarial learning of object-aware activation map for weakly-supervised semantic segmentation," *TCSVT*, 2023.

[20] F. Meng, K. Luo, H. Li, Q. Wu, and X. Xu, "Weakly supervised semantic segmentation by a class-level multiple group cosegmentation and foreground fusion strategy," *TCSVT*, vol. 30, no. 12, pp. 4823–4836, 2019.

[21] Z. Qin, Y. Chen, G. Zhu, E. Zhou, Y. Zhou, Y. Zhou, and C. Zhu, "Enhanced pseudo-label generation with self-supervised training for weakly-supervised semantic segmentation," *TCSVT*, 2024.

[22] X. Chang, H. Pan, W. Sun, and H. Gao, "A multi-phase camera-lidar fusion network for 3d semantic segmentation with weak supervision," *TCSVT*, 2023.

[23] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *TPAMI*, vol. 24, no. 5, pp. 603–619, 2002.

[24] K. Li, Z. Wang, Z. Cheng, R. Yu, Y. Zhao, G. Song, C. Liu, L. Yuan, and J. Chen, "Acseg: Adaptive conceptualization for unsupervised semantic segmentation," in *CVPR*, 2023, pp. 7162–7172.

[25] Z. Zhang, B. Yang, B. Wang, and B. Li, "Growsp: Unsupervised semantic segmentation of 3d point clouds," in *CVPR*, 2023, pp. 17 619–17 629.

[26] Y. Cao, H. Zhang, X. Lu, Y. Chen, Z. Xiao, and Y. Wang, "Adaptive refining-aggregation-separation framework for unsupervised domain adaptation semantic segmentation," *TCSVT*, 2023.

[27] H. S. Seong, W. Moon, S. Lee, and J.-P. Heo, "Leveraging hidden positives for unsupervised semantic segmentation," in *CVPR*, 2023, pp. 19 540–19 549.

[28] J. Wu, H. Fan, Z. Li, G.-H. Liu, and S. Lin, "Information transfer in semi-supervised semantic segmentation," *TCSVT*, 2023.

[29] X. Lu, L. Jiao, L. Li, F. Liu, X. Liu, and S. Yang, "Self pseudo entropy knowledge distillation for semi-supervised semantic segmentation," *TCSVT*, 2024.

[30] X. Zhang, Y. Yang, L. Ran, L. Chen, K. Wang, L. Yu, P. Wang, and Y. Zhang, "Remote sensing image semantic change detection boosted by semi-supervised contrastive learning of semantic segmentation," *TGRS*, vol. 62, pp. 1–13, 2024.

[31] J. Kim, K. Ryoo, J. Seo, G. Lee, D. Kim, H. Cho, and S. Kim, "Semi-supervised learning of semantic correspondence with pseudo-labels," in *CVPR*, 2022, pp. 19 699–19 709.

[32] Y. Zhang, Z. Gong, X. Zhao, X. Zheng, and W. Yao, "Semi-supervised semantic segmentation with uncertainty-guided self cross supervision," in *ACCV*, 2022, pp. 4631–4647.

[33] L. Yang, W. Zhuo, L. Qi, Y. Shi, and Y. Gao, "St++: Make self-training work better for semi-supervised semantic segmentation," in *CVPR*, 2022, pp. 4268–4277.

[34] Y. Ouali, C. Hudelot, and M. Tami, "Semi-supervised semantic segmentation with cross-consistency training," in *CVPR*, 2020, pp. 12 674–12 684.

[35] Z. Ke, D. Qiu, K. Li, Q. Yan, and R. W. Lau, "Guided collaborative training for pixel-wise semi-supervised learning," in *ECCV*. Springer, 2020, pp. 429–445.

[36] J. Lee, E. Kim, and S. Yoon, "Anti-adversarially manipulated attributions for weakly and semi-supervised semantic segmentation," in *CVPR*, 2021, pp. 4071–4080.

[37] N. Souly, C. Spampinato, and M. Shah, "Semi supervised semantic segmentation using generative adversarial network," in *ICCV*, 2017, pp. 5688–5696.

[38] 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 *CVPR*, 2021, pp. 8300–8311.

[39] I. Alonso, A. Sabater, D. Ferstl, L. Montesano, and A. C. Murillo, "Semi-supervised semantic segmentation with pixel-level contrastive learning from a class-wise memory bank," in *ICCV*, 2021, pp. 8219–8228.

[40] S. Liu, S. Zhi, E. Johns, and A. J. Davison, "Bootstrapping semantic segmentation with regional contrast," in *ICLR*, 2022.

[41] P. Qiao, Z. Wei, Y. Wang, Z. Wang, G. Song, F. Xu, X. Ji, C. Liu, and J. Chen, "Fuzzy positive learning for semi-supervised semantic segmentation," in *CVPR*, 2023, pp. 15 465–15 474.

[42] C. Wang, H. Xie, Y. Yuan, C. Fu, and X. Yue, "Space engage: Collaborative space supervision for contrastive-based semi-supervised semantic segmentation," in *ICCV*, 2023, pp. 931–942.

[43] 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 *CVPR*, 2021, pp. 1205–1214.

[44] J. Zhang, T. Wu, C. Ding, H. Zhao, and G. Guo, "Region-level contrastive and consistency learning for semi-supervised semantic segmentation," in *IJCAI*, 2022, pp. 1622–1628.

[45] H. Xiao, D. Li, H. Xu, S. Fu, D. Yan, K. Song, and C. Peng, "Semi-supervised semantic segmentation with cross teacher training," *Neurocomputing*, vol. 508, pp. 36–46, 2022.

[46] D.-H. Lee *et al.*, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *ICML*, vol. 3, no. 2. Atlanta, 2013, p. 896.

[47] M. Seydgar, S. Rahnamayan, P. Ghamisi, and A. A. Bidgoli, "Semisupervised hyperspectral image classification using a probabilistic pseudo-label generation framework," *TGRS*, vol. 60, pp. 1–18, 2022.

[48] S. Hu, C.-H. Liu, J. Dutta, M.-C. Chang, S. Lyu, and N. Ramakrishnan, "Pseudoprop: Robust pseudo-label generation for semi-supervised object detection in autonomous driving systems," in *CVPR*, 2022, pp. 4390–4398.

[49] X. Zhang, Y. Yang, L. Ran, L. Chen, K. Wang, L. Yu, P. Wang, and Y. Zhang, "Remote sensing image semantic change detection boosted by semi-supervised contrastive learning of semantic segmentation," *TGRS*, vol. 62, pp. 1–13, 2024.

[50] X. Huo, L. Xie, J. He, Z. Yang, W. Zhou, H. Li, and Q. Tian, "Atso: Asynchronous teacher-student optimization for semi-supervised image segmentation," in *CVPR*, 2021, pp. 1235–1244.

[51] L. Li, W. Zhang, X. Zhang, M. Emam, and W. Jing, "Semi-supervised remote sensing image semantic segmentation method based on deep learning," *Electronics*, vol. 12, no. 2, p. 348, 2023.

[52] J. Xu, Y. Jiang, B. Yuan, S. Li, and T. Song, "Automated scoring of clinical patient notes using advanced nlp and pseudo labeling," in *ICAICA*. IEEE, 2023, pp. 384–388.

[53] Z. Chen, Y. Luo, Z. Wang, M. Baktashmotlagh, and Z. Huang, "Revisiting domain-adaptive 3d object detection by reliable, diverse and class-balanced pseudo-labeling," in *ICCV*, 2023, pp. 3714–3726.

[54] Y. Higuchi, N. Moritz, J. L. Roux, and T. Hori, "Momentum pseudo-labeling for semi-supervised speech recognition," *arXiv:2106.08922*, 2021.

[55] Y. Xu, F. Wei, X. Sun, C. Yang, Y. Shen, B. Dai, B. Zhou, and S. Lin, "Cross-model pseudo-labeling for semi-supervised action recognition," in *CVPR*, 2022, pp. 2959–2968.

[56] J. E. van Engelen and H. H. Hoos, "A survey on semi-supervised learning," *Machine Learning*, vol. 109, no. 2, pp. 373–440, Feb. 2020.

[57] A. Peláez-Vegas, P. Mesejo, and J. Luengo, "A survey on semi-supervised semantic segmentation," *arXiv:2302.09899*, 2023.

[58] 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.

[59] Q. Xie, M.-T. Luong, E. Hovy, and Q. V. Le, "Self-training with noisy student improves imagenet classification," in *CVPR*, 2020, pp. 10 687–10 698.

[60] E. Arazo, D. Ortego, P. Albert, N. E. O'Connor, and K. McGuinness, "Pseudo-labeling and confirmation bias in deep semi-supervised learning," in *IJCNN*. IEEE, 2020, pp. 1–8.

[61] J. Liu, C. Desrosiers, and Y. Zhou, "Semi-supervised medical image segmentation using cross-model pseudo-supervision with shape awareness and local context constraints," in *MICCAI*. Springer, 2022, pp. 140–150.

[62] M.-C. Xu, Y. Zhou, C. Jin, M. de Groot, D. C. Alexander, N. P. Oxtoby, Y. Hu, and J. Jacob, "Expectation maximization pseudo labelling for segmentation with limited annotations," *arXiv:2305.01747*, 2023.

[63] X.-Y. Tong, G.-S. Xia, Q. Lu, H. Shen, S. Li, S. You, and L. Zhang, "Land-cover classification with high-resolution remote sensing images using transferable deep models," *Remote Sensing of Environment*, vol. 237, p. 111322, 2020.

[64] J.-X. Wang, S.-B. Chen, C. H. Ding, J. Tang, and B. Luo, "Semi-supervised semantic segmentation of remote sensing images with iterative contrastive network," *IEEE GRSL*, vol. 19, pp. 1–5, 2022.

[65] Z. Li, B. Ko, and H.-J. Choi, "Naive semi-supervised deep learning using pseudo-label," *PEER PEER NETW APPL*, vol. 12, pp. 1358–1368, 2019.

[66] M. N. Rizve, K. Duarte, Y. S. Rawat, and M. Shah, "In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning," *arXiv:2101.06329*, 2021.

[67] P. Cascante-Bonilla, F. Tan, Y. Qi, and V. Ordonez, "Curriculum labeling: Revisiting pseudo-labeling for semi-supervised learning," in *AAAI*, vol. 35, no. 8, 2021, pp. 6912–6920.

[68] M. S. Ibrahim, A. Vahdat, M. Ranjbar, and W. G. Macready, "Semi-supervised semantic image segmentation with self-correcting networks," in *CVPR*, 2020, pp. 12 715–12 725.

[69] J. Ma, C. Wang, Y. Liu, L. Lin, and G. Li, "Enhanced soft label for semi-supervised semantic segmentation," in *ICCV*, 2023, pp. 1185–1195.

[70] S. Mittal, M. Tatarchenko, and T. Brox, "Semi-supervised semantic segmentation with high-and low-level consistency," *TPAMI*, vol. 43, no. 4, pp. 1369–1379, 2019.

[71] B. Sun, Y. Yang, L. Zhang, M.-M. Cheng, and Q. Hou, "Corrmatch: Label propagation via correlation matching for semi-supervised semantic segmentation," in *CVPR*, 2024, pp. 3097–3107.

[72] 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:2010.09713*, 2020.

[73] H. Wu, X. Li, Y. Lin, and K.-T. Cheng, "Compete to win: Enhancing pseudo labels for barely-supervised medical image segmentation," *TMI*, 2023.

[74] 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," in *NeurIPS*, vol. 33, 2020, pp. 596–608.

[75] L. Ran, W. Zhan, Y. Li, X. Zhang, and S. Zhang, "Dtfseg: A dynamic threshold filtering method for semi-supervised semantic segmentation," in *China Automation Congress (CAC)*. IEEE, 2023, pp. 7571–7576.

[76] Z. Feng, Q. Zhou, G. Cheng, X. Tan, J. Shi, and L. Ma, "Semi-supervised semantic segmentation via dynamic self-training and class-balanced curriculum," *arXiv:2004.08514*, 2020.

[77] R. Ke, A. I. Aviles-Rivero, S. Pandey, S. Reddy, and C.-B. Schönlieb, "A three-stage self-training framework for semi-supervised semantic segmentation," *TIP*, vol. 31, pp. 1805–1815, 2022.

[78] R. Yi, Y. Huang, Q. Guan, M. Pu, and R. Zhang, "Learning from pixel-level label noise: A new perspective for semi-supervised semantic segmentation," *TIP*, vol. 31, pp. 623–635, 2021.

[79] X. Wang, J. Xiao, B. Zhang, and L. Yu, "Card: Semi-supervised semantic segmentation via class-agnostic relation based denoising." in *IJCAI*, 2022, pp. 1451–1457.

[80] Z. Li, B. Ko, and H. Choi, "Pseudo-labeling using gaussian process for semi-supervised deep learning," in *International Conference on Big Data and Smart Computing (BigComp)*. IEEE, 2018, pp. 263–269.

[81] D. Yarowsky, "Unsupervised word sense disambiguation rivaling supervised methods," in *ACL*, 1995, pp. 189–196.

[82] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4*. Springer, 2018, pp. 3–11.

[83] T. Miyato, S.-i. Maeda, M. Koyama, and S. Ishii, "Virtual adversarial training: a regularization method for supervised and semi-supervised learning," *TPAMI*, vol. 41, no. 8, pp. 1979–1993, 2018.

[84] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," *NeurIPS*, vol. 30, 2017.

[85] M. Xu, Z. Zhang, H. Hu, J. Wang, L. Wang, F. Wei, X. Bai, and Z. Liu, "End-to-end semi-supervised object detection with soft teacher," in *ICCV*. Montreal, QC, Canada: IEEE, Oct. 2021, pp. 3040–3049.

[86] 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 *CRV*. IEEE, 2022, pp. 58–66.

[87] 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," *TPAMI*, vol. 40, no. 4, pp. 834–848, 2017.

[88] L. Wu, L. Fang, X. He, M. He, J. Ma, and Z. Zhong, "Querying labeled for unlabeled: Cross-image semantic consistency guided semi-supervised semantic segmentation," *TPAMI*, 2023.

[89] X. Chen, Y. Yuan, G. Zeng, and J. Wang, "Semi-supervised semantic segmentation with cross pseudo supervision," in *CVPR*, 2021, pp. 2613–2622.

[90] H. Li and H. Zheng, "A residual correction approach for semi-supervised semantic segmentation," in *PRCV 2021*. Springer, 2021, pp. 90–102.

[91] D. Kwon and S. Kwak, "Semi-supervised semantic segmentation with error localization network," in *CVPR*, 2022, pp. 9957–9967.

[92] R. Mendel, L. A. De Souza, D. Rauber, J. P. Papa, and C. Palm, "Semi-supervised segmentation based on error-correcting supervision," in *ECCV*. Springer, 2020, pp. 141–157.

[93] M. Lee, S. Lee, J. Lee, and H. Shim, "Saliency as pseudo-pixel supervision for weakly and semi-supervised semantic segmentation," *TPAMI*, 2023.

[94] P. Tu, Y. Huang, F. Zheng, Z. He, L. Cao, and L. Shao, "Guidedmix-net: Semi-supervised semantic segmentation by using labeled images as reference," in *AAAI*, vol. 36, no. 2, 2022, pp. 2379–2387.

[95] H. Wang, Q. Zhang, Y. Li, and X. Li, "Allspark: Reborn labeled features from unlabeled in transformer for semi-supervised semantic segmentation," in *CVPR*, 2024, pp. 3627–3636.

[96] 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.

[97] J. Ju, H. Noh, Y. Wang, M. Seo, and D.-G. Choi, "Cafs: Class adaptive framework for semi-supervised semantic segmentation," *arXiv:2303.11606*, 2023.

[98] D. Filipiak, P. Tempczyk, and M. Cygan, "n-cps: Generalising cross pseudo supervision to n networks for semi-supervised semantic segmentation," *arXiv:2112.07528*, 2021.

[99] J. Fan, B. Gao, H. Jin, and L. Jiang, "Ucc: Uncertainty guided cross-head co-training for semi-supervised semantic segmentation," in *CVPR*, 2022, pp. 9947–9956.

[100] Z. Wang, Z. Zhao, X. Xing, D. Xu, X. Kong, and L. Zhou, "Conflict-based cross-view consistency for semi-supervised semantic segmentation," in *CVPR*, 2023, pp. 19 585–19 595.

[101] C. Liang, W. Wang, J. Miao, and Y. Yang, "Logic-induced diagnostic reasoning for semi-supervised semantic segmentation," in *ICCV*, 2023, pp. 16 197–16 208.

[102] R. He, J. Yang, and X. Qi, "Re-distributing biased pseudo labels for semi-supervised semantic segmentation: A baseline investigation," in *ICCV*, 2021, pp. 6930–6940.

[103] B. Chen, J. Jiang, X. Wang, P. Wan, J. Wang, and M. Long, "Debiased self-training for semi-supervised learning," *NeurIPS*, vol. 35, pp. 32 424–32 437, 2022.

[104] Y. Zhou, H. Xu, W. Zhang, B. Gao, and P.-A. Heng, "C3-semiseg: Contrastive semi-supervised segmentation via cross-set learning and dynamic class-balancing," in *ICCV*, 2021, pp. 7036–7045.

[105] R. Chen, T. Chen, Q. Wang, and Y. Yao, "Semi-supervised semantic segmentation with region relevance," *arXiv:2304.11539*, 2023.

[106] H. Kong, G.-H. Lee, S. Kim, and S.-W. Lee, "Pruning-guided curriculum learning for semi-supervised semantic segmentation," in *WACV*, 2023, pp. 5914–5923.

[107] Y. Jin, J. Wang, and D. Lin, "Semi-supervised semantic segmentation via gentle teaching assistant," *NeurIPS*, vol. 35, pp. 2803–2816, 2022.

[108] 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 *ICCV*, 2019, pp. 6023–6032.

[109] V. Olsson, W. Tranheden, J. Pinto, and L. Svensson, "Classmix: Segmentation-based data augmentation for semi-supervised learning," in *WACV*, 2021, pp. 1369–1378.

[110] Y. Chen, X. Ouyang, K. Zhu, and G. Agam, "Complexmix: Semi-supervised semantic segmentation via mask-based data augmentation," in *ICIP*. IEEE, 2021, pp. 2264–2268.

[111] J. Yuan, Y. Liu, C. Shen, Z. Wang, and H. Li, "A simple baseline for semi-supervised semantic segmentation with strong data augmentation," in *ICCV*, 2021, pp. 8229–8238.

[112] Z. Zhao, L. Yang, S. Long, J. Pi, L. Zhou, and J. Wang, "Augmentation matters: A simple-yet-effective approach to semi-supervised semantic segmentation," in *CVPR*, 2023, pp. 11 350–11 359.

[113] Z. Zhao, S. Long, J. Pi, J. Wang, and L. Zhou, "Instance-specific and model-adaptive supervision for semi-supervised semantic segmentation," in *CVPR*, 2023, pp. 23 705–23 714.

[114] I. Grubišić, M. Oršić, and S. Šegvić, "A baseline for semi-supervised learning of efficient semantic segmentation models," in *MVA*. IEEE, 2021, pp. 1–5.

[115] C. Cao, T. Lin, D. He, F. Li, H. Yue, J. Yang, and E. Ding, "Adversarial dual-student with differentiable spatial warping for semi-supervised semantic segmentation," *TCSVT*, vol. 33, no. 2, pp. 793–803, 2022.

[116] I. Grubišić, M. Oršić, and S. Šegvić, "Revisiting consistency for semi-supervised semantic segmentation," *Sensors*, vol. 23, no. 2, p. 940, 2023.

[117] Y. Wang, J. Zhang, M. Kan, and S. Shan, "Learning pseudo labels for semi-and-weakly supervised semantic segmentation," *Pattern Recognition*, vol. 132, p. 108925, 2022.

[118] S. Fan, F. Zhu, Z. Feng, Y. Lv, M. Song, and F.-Y. Wang, "Conservative-progressive collaborative learning for semi-supervised semantic segmentation," *TIP*, 2023.

[119] Y. Zhang, T. Xiang, T. M. Hospedales, and H. Lu, "Deep mutual learning," in *CVPR*, 2018, pp. 4320–4328.

[120] W. Wang and Z.-H. Zhou, "A new analysis of co-training." in *ICML*, vol. 2, 2010, p. 3.

[121] J. Peng, G. Estrada, M. Pedersoli, and C. Desrosiers, "Deep co-training for semi-supervised image segmentation," *Pattern Recognition*, vol. 107, p. 107269, 2020.

[122] Y. Xia, D. Yang, Z. Yu, F. Liu, J. Cai, L. Yu, Z. Zhu, D. Xu, A. Yuille, and H. Roth, "Uncertainty-aware multi-view co-training for semi-supervised medical image segmentation and domain adaptation," *Medical image analysis*, vol. 65, p. 101766, 2020.

[123] W. Zhang, W. Ouyang, W. Li, and D. Xu, "Collaborative and adversarial network for unsupervised domain adaptation," in *CVPR*, 2018.

[124] L. Yang, L. Qi, L. Feng, W. Zhang, and Y. Shi, "Revisiting weak-to-strong consistency in semi-supervised semantic segmentation," in *CVPR*, 2023, pp. 7236–7246.

[125] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *ICML*, 2009, pp. 41–48.

[126] M. McCloskey and N. J. Cohen, "Catastrophic interference in connectionist networks: The sequential learning problem," in *Psychology of learning and motivation*. Elsevier, 1989, vol. 24, pp. 109–165.

[127] 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.

[128] Z. Cai, X. Yan, Y. Wu, K. Ma, J. Cheng, and F. Yu, "Dgcl: an efficient communication library for distributed gnn training," in *EuroSys*, 2021, pp. 130–144.

[129] 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:2205.07246*, 2022.

[130] E. Diao, J. Ding, and V. Tarokh, "Semifl: Semi-supervised federated learning for unlabeled clients with alternate training," *NeurIPS*, vol. 35, pp. 17 871–17 884, 2022.

[131] Q. Xie, Z. Dai, E. Hovy, T. Luong, and Q. Le, "Unsupervised data augmentation for consistency training," *NeurIPS*, vol. 33, pp. 6256–6268, 2020.

[132] Y. Zou, Z. Yu, B. Kumar, and J. Wang, "Unsupervised domain adaptation for semantic segmentation via class-balanced self-training," in *ECCV*, 2018, pp. 289–305.

[133] Y. Liu, Y. Tian, Y. Chen, F. Liu, V. Belagiannis, and G. Carneiro, "Perturbed and strict mean teachers for semi-supervised semantic segmentation," in *CVPR*, 2022, pp. 4258–4267.

[134] Z. Ke, D. Wang, Q. Yan, J. Ren, and R. W. Lau, "Dual student: Breaking the limits of the teacher in semi-supervised learning," in *ICCV*, 2019, pp. 6728–6736.

[135] 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," *NeurIPS*, vol. 33, pp. 596–608, 2020.

[136] G. French, S. Laine, T. Aila, M. Mackiewicz, and G. Finlayson, "Semi-supervised semantic segmentation needs strong, varied perturbations," *arXiv:1906.01916*, 2019.

[137] Y. Zhong, B. Yuan, H. Wu, Z. Yuan, J. Peng, and Y.-X. Wang, "Pixel contrastive-consistent semi-supervised semantic segmentation," in *ICCV*, 2021, pp. 7273–7282.

[138] S. Li, Y. He, W. Zhang, W. Zhang, X. Tan, J. Han, E. Ding, and J. Wang, "Cfcg: Semi-supervised semantic segmentation via cross-fusion and contour guidance supervision," in *ICCV*, 2023, pp. 16 348–16 358.

[139] X. Wang, B. Zhang, L. Yu, and J. Xiao, "Hunting sparsity: Density-guided contrastive learning for semi-supervised semantic segmentation," in *CVPR*, 2023, pp. 3114–3123.

[140] P. Bachman, O. Alsharif, and D. Precup, "Learning with pseudo-ensembles," *NeurIPS*, vol. 27, 2014.

[141] Q. Li, Y. Shi, and X. X. Zhu, "Semi-supervised building footprint generation with feature and output consistency training," *TGRS*, vol. 60, pp. 1–17, 2022.

[142] C. Nong, X. Fan, and J. Wang, "Semi-supervised learning for weed and crop segmentation using uav imagery," *Frontiers in Plant Science*, vol. 13, p. 927368, 2022.

[143] R. M. S. Bashir, T. Qaiser, S. E. A. Raza, and N. M. Rajpoot, "Consistency regularisation in varying contexts and feature perturbations for semi-supervised semantic segmentation of histology images," *arXiv:2301.13141*, 2023.

[144] S. An, H. Zhu, J. Zhang, J. Ye, S. Wang, J. Yin, and H. Zhang, "Deep tri-training for semi-supervised image segmentation," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 10 097–10 104, 2022.

[145] Y. He, J. Wang, C. Liao, B. Shan, and X. Zhou, "Classhyper: Classmix-based hybrid perturbations for deep semi-supervised semantic segmentation of remote sensing imagery," *Remote Sensing*, vol. 14, no. 4, p. 879, 2022.

[146] Y. Wu, C. Liu, L. Chen, D. Zhao, Q. Zheng, and H. Zhou, "Perturbation consistency and mutual information regularization for semi-supervised semantic segmentation," *Multimedia Systems*, vol. 29, no. 2, pp. 511–523, 2023.

[147] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, and C. A. Raffel, "Mixmatch: A holistic approach to semi-supervised learning," *NeurIPS*, vol. 32, 2019.

[148] J. Yuan, J. Ge, Q. Qian, Z. Wang, F. Wang, and Y. Liu, "Semi-supervised semantic segmentation with mutual knowledge distillation," *arXiv:2208.11499*, 2022.

[149] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," *NeurIPS*, vol. 33, pp. 9912–9924, 2020.

[150] D. Berthelot, N. Carlini, E. D. Cubuk, A. Kurakin, K. Sohn, H. Zhang, and C. Raffel, "Remixmatch: Semi-supervised learning with distribution matching and augmentation anchoring," in *ICLR*, 2020.

[151] H. R. Roth, L. Lu, A. Farag, H.-C. Shin, J. Liu, E. B. Turkbey, and R. M. Summers, "Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation," in *MICCAI*. Springer, 2015, pp. 556–564.

[152] O. Bernard, A. Lalande, C. Zotti, F. Cervenansky, X. Yang, P.-A. Heng, I. Cetin, K. Lekadir, O. Camara, M. A. G. Ballester *et al.*, "Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved?" *TMI*, vol. 37, no. 11, pp. 2514–2525, 2018.

[153] G. Litjens, R. Toth, W. Van De Ven, C. Hoeks, S. Kerkstra, B. Van Ginneken, G. Vincent, G. Guillard, N. Birbeck, J. Zhang *et al.*, "Evaluation of prostate segmentation algorithms for mri: the promise12 challenge," *Medical image analysis*, vol. 18, no. 2, pp. 359–373, 2014.

[154] A. L. Simpson, M. Antonelli, S. Bakas, M. Bilello, K. Farahani, B. Van Ginneken, A. Kopp-Schneider, B. A. Landman, G. Litjens, B. Menze *et al.*, "A large annotated medical image dataset for the development and evaluation of segmentation algorithms," *arXiv:1902.09063*, 2019.

[155] J.-P. Charbonnier, M. Brink, F. Ciompi, E. T. Scholten, C. M. Schaefer-Prokop, and E. M. Van Rikxoort, "Automatic pulmonary artery-vein separation and classification in computed tomography using tree partitioning and peripheral vessel matching," *TMI*, vol. 35, no. 3, pp. 882–892, 2015.

[156] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest *et al.*, "The multimodal brain tumor image segmentation benchmark (brats)," *TMI*, vol. 34, no. 10, pp. 1993–2024, 2014.

[157] M. Antonelli, A. Reinke, S. Bakas, K. Farahani, A. Kopp-Schneider, B. A. Landman, G. Litjens, B. Menze, O. Ronneberger, R. M. Summers *et al.*, "The medical segmentation decathlon," *Nature communications*, vol. 13, no. 1, p. 4128, 2022.

[158] Y. Bai, D. Chen, Q. Li, W. Shen, and Y. Wang, "Bidirectional copy-paste for semi-supervised medical image segmentation," in *CVPR*, 2023, pp. 11 514–11 524.

[159] Z. Xiong, Q. Xia, Z. Hu, N. Huang, C. Bian, Y. Zheng, S. Vesal, N. Ravikumar, A. Maier, X. Yang *et al.*, "A global benchmark of algorithms for segmenting the left atrium from late gadolinium-enhanced cardiac magnetic resonance imaging," *Medical image analysis*, vol. 67, p. 101832, 2021.

[160] H. Yao, X. Hu, and X. Li, "Enhancing pseudo label quality for semi-supervised domain-generalized medical image segmentation," in *AAAI*, vol. 36, no. 3, 2022, pp. 3099–3107.

[161] V. M. Campello, P. Gkontra, C. Izquierdo, C. Martin-Isla, A. Sojoudi, P. M. Full, K. Maier-Hein, Y. Zhang, Z. He, J. Ma *et al.*, "Multi-centre, multi-vendor and multi-disease cardiac segmentation: the m&ms challenge," *TMI*, vol. 40, no. 12, pp. 3543–3554, 2021.

[162] F. Prados, J. Ashburner, C. Blaiotta, T. Brosch, J. Carballido-Gamio, M. J. Cardoso, B. N. Conrad, E. Datta, G. Dávid, B. De Leener *et al.*, "Spinal cord grey matter segmentation challenge," *Neuroimage*, vol. 152, pp. 312–329, 2017.

[163] T. Lei, D. Zhang, X. Du, X. Wang, Y. Wan, and A. K. Nandi, "Semi-supervised medical image segmentation using adversarial consistency learning and dynamic convolution network," *TMI*, 2022.

[164] P. Bilic, P. Christ, H. B. Li, E. Vorontsov, A. Ben-Cohen, G. Kaissis, A. Szeskin, C. Jacobs, G. E. H. Mamani, G. Chartrand *et al.*, "The liver tumor segmentation benchmark (lits)," *Medical Image Analysis*, vol. 84, p. 102680, 2023.

[165] N. Codella, V. Rotemberg, P. Tschandl, M. E. Celebi, S. Dusza, D. Gutman, B. Helba, A. Kalloo, K. Liopyris, M. Marchetti *et al.*, "Skin lesion analysis toward melanoma detection 2018: A challenge hosted by

the international skin imaging collaboration (isic)," *arXiv:1902.03368*, 2019.

[166] M. Tran, S. J. Wagner, M. Boxberg, and T. Peng, "S5cl: Unifying fully-supervised, self-supervised, and semi-supervised learning through hierarchical contrastive learning," in *MICCAI*. Springer, 2022, pp. 99–108.

[167] M. Macenko, M. Niethammer, J. S. Marron, D. Borland, J. T. Woosley, X. Guan, C. Schmitt, and N. E. Thomas, "A method for normalizing histology slides for quantitative analysis," in *ISBI*. IEEE, 2009, pp. 1107–1110.

[168] C. Matek, S. Schwarz, K. Spiekermann, and C. Marr, "Human-level recognition of blast cells in acute myeloid leukaemia with convolutional neural networks," *Nat. Mach. Intell.*, vol. 1, no. 11, pp. 538–544, 2019.

[169] X. Zhao, C. Fang, D.-J. Fan, X. Lin, F. Gao, and G. Li, "Cross-level contrastive learning and consistency constraint for semi-supervised medical image segmentation," in *ISBI*. IEEE, 2022, pp. 1–5.

[170] D. Jha, P. H. Smedsrud, M. A. Riegler, P. Halvorsen, T. de Lange, D. Johansen, and H. D. Johansen, "Kvasir-seg: A segmented polyp dataset," in *MMM*. Springer, 2020, pp. 451–462.

[171] H. Wu, Z. Wang, Y. Song, L. Yang, and J. Qin, "Cross-patch dense contrastive learning for semi-supervised segmentation of cellular nuclei in histopathologic images," in *CVPR*, 2022, pp. 11 666–11 675.

[172] J. C. Caicedo, A. Goodman, K. W. Karhohs, B. A. Cimini, J. Ackerman, M. Haghighi, C. Heng, T. Becker, M. Doan, C. McQuin *et al.*, "Nucleus segmentation across imaging experiments: the 2018 data science bowl," *Nature methods*, vol. 16, no. 12, pp. 1247–1253, 2019.

[173] N. Kumar, R. Verma, D. Anand, Y. Zhou, O. F. Onder, E. Tsougenis, H. Chen, P.-A. Heng, J. Li, Z. Hu *et al.*, "A multi-organ nucleus segmentation challenge," *TMI*, vol. 39, no. 5, pp. 1380–1391, 2019.

[174] F. Rottensteiner, G. Sohn, J. Jung, M. Gerke, C. Baillard, S. Benitez, and U. Breitkopf, "The isprs benchmark on urban object classification and 3d building reconstruction," *ISPRS Annals*, vol. 1, no. 1, pp. 293–298, 2012.

[175] M. Cui, K. Li, Y. Li, D. Kamuhanda, and C. J. Tessone, "Semi-supervised semantic segmentation of remote sensing images based on dual cross-entropy consistency," *Entropy*, vol. 25, no. 4, p. 681, 2023.

[176] S. Waqas Zamir, A. Arora, A. Gupta, S. Khan, G. Sun, F. Shahbaz Khan, F. Zhu, L. Shao, G.-S. Xia, and X. Bai, "isaid: A large-scale dataset for instance segmentation in aerial images," in *CVPR Workshops*, 2019, pp. 28–37.

[177] Y. He, J. Wang, C. Liao, X. Zhou, and B. Shan, "Ms4d-net: Multitask-based semi-supervised semantic segmentation framework with perturbed dual mean teachers for building damage assessment from high-resolution remote sensing imagery," *Remote Sensing*, vol. 15, no. 2, p. 478, 2023.

[178] R. Gupta, R. Hosfelt, S. Sajeev, N. Patel, B. Goodman, J. Doshi, E. Heim, H. Choset, and M. Gaston, "xbd: A dataset for assessing building damage from satellite imagery," *arXiv:1911.09296*, 2019.

[179] H. Hu, F. Wei, H. Hu, Q. Ye, J. Cui, and L. Wang, "Semi-supervised semantic segmentation via adaptive equalization learning," *NeurIPS*, vol. 34, pp. 22 106–22 118, 2021.

[180] D. Guan, J. Huang, A. Xiao, and S. Lu, "Unbiased subclass regularization for semi-supervised semantic segmentation," in *CVPR*, 2022, pp. 9968–9978.

[181] W. Lin, Z. He, and M. Xiao, "Balanced clustering: A uniform model and fast algorithm." in *IJCAI*, 2019, pp. 2987–2993.

[182] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv:1807.03748*, 2018.

[183] S. Liu, S. Zhi, E. Johns, and A. J. Davison, "Bootstrapping semantic segmentation with regional contrast," *arXiv:2104.04465*, 2021.

[184] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *NeurIPS*, vol. 27, 2014.

[185] W.-C. Hung, Y.-H. Tsai, Y.-T. Liou, Y.-Y. Lin, and M.-H. Yang, "Adversarial learning for semi-supervised semantic segmentation," *arXiv:1802.07934*, 2018.

[186] G. Jin, C. Liu, and X. Chen, "Adversarial network integrating dual attention and sparse representation for semi-supervised semantic segmentation," *IPMA*, vol. 58, no. 5, p. 102680, 2021.

[187] C. M. Bishop, "Pattern recognition and machine learning," *Springer google schola*, vol. 2, pp. 1122–1128, 2006.

[188] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *IJCV*, vol. 88, pp. 303–338, 2010.

[189] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, "Scene parsing through ade20k dataset," in *CVPR*, 2017, pp. 633–641.

[190] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *ECCV*. Springer, 2014, pp. 740–755.

[191] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *CVPR*. IEEE, 2012, pp. 3354–3361.

[192] G. J. Brostow, J. Fauqueur, and R. Cipolla, "Semantic object classes in video: A high-definition ground truth database," *Pattern recognition letters*, vol. 30, no. 2, pp. 88–97, 2009.

[193] X. Hu, C.-W. Fu, L. Zhu, and P.-A. Heng, "Depth-attentional features for single-image rain removal," in *CVPR*, 2019, pp. 8022–8031.

[194] M. Hahner, D. Dai, C. Sakaridis, J.-N. Zaech, and L. Van Gool, "Semantic understanding of foggy scenes with purely synthetic data," in *ITSC*. IEEE, 2019, pp. 3675–3681.

[195] F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, and T. Darrell, "Bdd100k: A diverse driving dataset for heterogeneous multitask learning," in *CVPR*, 2020, pp. 2636–2645.

[196] G. Neuhold, T. Ollmann, S. Rota Bulo, and P. Kontschieder, "The mapillary vistas dataset for semantic understanding of street scenes," in *ICCV*, 2017, pp. 4990–4999.

[197] N. Kumar, R. Verma, S. Sharma, S. Bhargava, A. Vahadane, and A. Sethi, "A dataset and a technique for generalized nuclear segmentation for computational pathology," *TMI*, vol. 36, no. 7, pp. 1550–1560, 2017.

[198] M. Rahnemoonfar, T. Chowdhury, A. Sarkar, D. Varshney, M. Yari, and R. R. Murphy, "Floodnet: A high resolution aerial imagery dataset for post flood scene understanding," *IEEE Access*, vol. 9, pp. 89 644–89 654, 2021.

[199] P. Helber, B. Bischke, A. Dengel, and D. Borth, "Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification," *IEEE J-STARS*, vol. 12, no. 7, pp. 2217–2226, 2019.

[200] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention*. Springer, 2015, pp. 234–241.

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

[202] D. Lu and Q. Weng, "A survey of image classification methods and techniques for improving classification performance," *IJRS*, vol. 28, no. 5, pp. 823–870, 2007.

[203] L. Ma, P. Peng, G. Chen, Y. Zhao, S. Dong, and Y. Tian, "Picking up quantization steps for compressed image classification," *TCSVT*, vol. 33, no. 4, pp. 1884–1898, 2022.

[204] X. Li, Q. Song, J. Wu, R. Zhu, Z. Ma, and J.-H. Xue, "Locally-enriched cross-reconstruction for few-shot fine-grained image classification," *TCSVT*, 2023.

[205] M. Ma, W. Ma, L. Jiao, X. Liu, F. Liu, L. Li, and S. Yang, "Mbsi-net: Multimodal balanced self-learning interaction network for image classification," *TCSVT*, 2023.

[206] H. Hu, J. Gu, Z. Zhang, J. Dai, and Y. Wei, "Relation networks for object detection," in *CVPR*, 2018, pp. 3588–3597.

[207] Z. Zou, K. Chen, Z. Shi, Y. Guo, and J. Ye, "Object detection in 20 years: A survey," *IEEE*, vol. 111, no. 3, pp. 257–276, 2023.

[208] Y. Tang, Z. Cao, Y. Yang, J. Liu, and J. Yu, "Semi-supervised few-shot object detection via adaptive pseudo labeling," *TCSVT*, 2023.

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

[210] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *CVPR*. IEEE, 2009, pp. 248–255.

[211] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *ICCV*, 2021, pp. 10 012–10 022.

[212] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," in *ICCV*, 2023, pp. 4015–4026.

[213] X. Yang and X. Gong, "Foundation model assisted weakly supervised semantic segmentation," in *WACV*, 2024, pp. 523–532.

[214] X. Wang, J. Yang, and T. Darrell, "Segment anything without supervision," *arXiv:2406.20081*, 2024.

[215] D. Yang, J. Ji, Y. Ma, T. Guo, H. Wang, X. Sun, and R. Ji, "Sam as the guide: Mastering pseudo-label refinement in semi-supervised referring expression segmentation," *arXiv:2406.01451*, 2024.

[216] C. Lin, Y. Liu, D. Wang, and L. Lin, "S-acmt: Enhancing semi-supervised medical image segmentation with sam as a cross-supervised secondary teacher," in *ICSIP*. IEEE, 2024, pp. 774–779.

[217] H. Huang, L. Lin, Y. Zhang, Y. Xu, J. Zheng, X. Mao, X. Qian, Z. Peng, J. Zhou, Y.-W. Chen *et al.*, "Graph-bas3net: Boundary-aware semi-supervised segmentation network with bilateral graph convolution," in *ICCV*, 2021, pp. 7386–7395.
[218] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *TPAMI*, vol. 41, no. 2, pp. 423–443, 2018.
[219] K. Wang, Q. Yin, W. Wang, S. Wu, and L. Wang, "A comprehensive survey on cross-modal retrieval," *arXiv:1607.06215*, 2016.
[220] Q. Cheng, Z. Tan, K. Wen, C. Chen, and X. Gu, "Semantic pre-alignment and ranking learning with unified framework for cross-modal retrieval," *TCSVT*, 2022.
[221] Y. Chen, D. Li, P. Zhang, J. Sui, Q. Lv, L. Tun, and L. Shang, "Cross-modal ambiguity learning for multimodal fake news detection," in *WWW*, 2022, pp. 2897–2905.
[222] R. Yadav, A. Samir, H. Rashed, S. Yogamani, and R. Dahyot, "Cnn based color and thermal image fusion for object detection in automated driving," *IMVIP*, vol. 2, 2020.
[223] M. Zhou, L. Zhou, S. Wang, Y. Cheng, L. Li, Z. Yu, and J. Liu, "Uc2: Universal cross-lingual cross-modal vision-and-language pre-training," in *CVPR*, 2021, pp. 4155–4165.
[224] L. Kirsch, J. Kunze, and D. Barber, "Modular networks: Learning to decompose neural computation," *NeurIPS*, vol. 31, 2018.
[225] J. Yang, Y. Huang, K. Niu, L. Huang, Z. Ma, and L. Wang, "Actor and action modular network for text-based video segmentation," *TIP*, vol. 31, pp. 4474–4489, 2022.
[226] X. Liang, "Learning personalized modular network guided by structured knowledge," in *CVPR*, 2019, pp. 8944–8952.
[227] P. K. Atrey, M. A. Hossain, A. El Saddik, and M. S. Kankanhalli, "Multimodal fusion for multimedia analysis: a survey," *Multimedia systems*, vol. 16, pp. 345–379, 2010.
[228] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *ICML*. PMLR, 2021, pp. 8748–8763.
[229] A. Farahani, S. Voghoei, K. Rasheed, and H. R. Arabnia, "A brief review of domain adaptation," *ICDATA*, pp. 877–894, 2021.
[230] U. Michieli, M. Biasetton, G. Agresti, and P. Zanuttigh, "Adversarial learning and self-teaching techniques for domain adaptation in semantic segmentation," *IEEE T-IV*, vol. 5, no. 3, pp. 508–518, 2020.
[231] D. Alexey, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv: 2010.11929*, 2020.
[232] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *CVPR*, 2022, pp. 16 000–16 009.
[233] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *ICCV*, 2021, pp. 9650–9660.
[234] B. Shi, X. Zhang, H. Xu, W. Dai, J. Zou, H. Xiong, and Q. Tian, "Multi-dataset pretraining: A unified model for semantic segmentation," *arXiv:2106.04121*, 2021.
[235] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," *IJCV*, vol. 130, no. 9, pp. 2337–2348, 2022.
[236] C. Ge, R. Huang, M. Xie, Z. Lai, S. Song, S. Li, and G. Huang, "Domain adaptation via prompt learning," *TNNLS*, 2023.
[237] M. U. Khattak, H. Rasheed, M. Maaz, S. Khan, and F. S. Khan, "Maple: Multi-modal prompt learning," in *CVPR*, 2023, pp. 19 113–19 122.
[238] B. Zhou, H. Zhao, X. Puig, T. Xiao, S. Fidler, A. Barriuso, and A. Torralba, "Semantic understanding of scenes through the ade20k dataset," *International Journal of Computer Vision*, vol. 127, pp. 302–321, 2019.
[239] S. Vandenhende, S. Georgoulis, W. Van Gansbeke, M. Proesmans, D. Dai, and L. Van Gool, "Multi-task learning for dense prediction tasks: A survey," *TPAMI*, vol. 44, no. 7, pp. 3614–3633, 2021.
[240] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *ICCV*, 2017, pp. 2961–2969.
[241] Y. Hu, R. Xian, Q. Wu, Q. Fan, L. Yin, and H. Zhao, "Revisiting scalarization in multi-task learning: A theoretical perspective," *NeurIPS*, vol. 36, 2024.
[242] K. Goel, P. Srinivasan, S. Tariq, and J. Philbin, "Quadronet: Multi-task learning for real-time semantic depth aware instance segmentation," in *WACV*, 2021, pp. 315–324.
[243] T. Xiao, Y. Liu, B. Zhou, Y. Jiang, and J. Sun, "Unified perceptual parsing for scene understanding," in *ECCV*, 2018, pp. 418–434.
[244] B. Settles, "Active learning literature survey," 2009.
[245] J. Yao, X. Cao, D. Hong, X. Wu, D. Meng, J. Chanussot, and Z. Xu, "Semi-active convolutional neural networks for hyperspectral image classification," *TGRS*, vol. 60, pp. 1–15, 2022.
[246] P. Mi, J. Lin, Y. Zhou, Y. Shen, G. Luo, X. Sun, L. Cao, R. Fu, Q. Xu, and R. Ji, "Active teacher for semi-supervised object detection," in *CVPR*, 2022, pp. 14 482–14 491.

Lingyan Ran received his B.S. and Ph.D. degrees from Northwestern Polytechnical University (NWPU), Xi'an, China, in 2011 and 2018. Earlier, he was a visiting scholar at Stevens Institute of Technology, Hoboken, NJ, from 2013 to 2015. He is currently an Associate Professor with the School of Computer Science, NWPU. His research interests include image classification, semantic segmentation, and change detection. He is currently a member of the China Computer Federation and China Society of Image and Graphics.

Yali Li received the B.Eng. degree from the Taiyuan University of Technology (TYUT), Taiyuan, China, in 2023, and is now pursuing her master's degree from Northwestern Polytechnical University (NWPU). Her main research interests are semantic segmentation and computer vision, etc.

Guoqiang Liang received the B.S. in automation and the Ph.D. degrees in pattern recognition and intelligent systems from Xi'an Jiaotong University, Xi'an, China in 2012 and 2018 respectively. From Aug. 2018 to Aug 2020, he did the Post-Doctoral Research at the School of Computer Science, Northwestern Polytechnical University (NWPU), Xi'an, China. Currently, he is an associate professor at NWPU. His research interests include continual learning and video summarization.

Yanning Zhang (Senior Member, IEEE) received the B.S. degree from Dalian University of Science and Engineering in 1988, and the M.S. and Ph.D. degrees from Northwestern Polytechnical University, Xi'an, China, in 1993 and 1996, respectively. She is a Professor with the School of Computer Science, at Northwestern Polytechnical University. She has published over 200 articles in international journals, conferences, and Chinese key journals. Her research interests include signal and image processing, computer vision, and pattern recognition.