Computer Vision and Image Understanding 249 (2024) 104146

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

Class Probability Space Regularization for semi-supervised semantic segmentation

Jianjian Yin [a], Shuai Yan [b], Tao Chen [b], Yi Chen [a], Yazhou Yao [b,*]

[a] School of Computer and Electronic Information/Artificial Intelligence, Nanjing Normal University, Nanjing, Jiangsu Province, China
[b] School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, Jiangsu Province, China

ARTICLE INFO

Keywords:
Human visual attention mechanisms
Semi-supervised semantic segmentation
Class probability space regularization
Class knowledge reshaping
Tail probability suppression

ABSTRACT

Semantic segmentation achieves fine-grained scene parsing in any scenario, making it one of the key research directions to facilitate the development of human visual attention mechanisms. Recent advancements in semi-supervised semantic segmentation have attracted considerable attention due to their potential in leveraging unlabeled data. However, existing methods only focus on exploring the knowledge of unlabeled pixels with high certainty prediction. Their insufficient mining of low certainty regions of unlabeled data results in a significant loss of supervisory information. Therefore, this paper proposes the **C**lass **P**robability **S**pace **R**egularization (**CPSR**) approach to further exploit the potential of each unlabeled pixel. Specifically, we first design a class knowledge reshaping module to regularize the probability space of low certainty pixels, thereby transforming them into high certainty ones for supervised training. Furthermore, we propose a tail probability suppression module to suppress the probabilities of tailed classes, which facilitates the network to learn more discriminative information from the class probability space. Extensive experiments conducted on the PASCAL VOC2012 and Cityscapes datasets prove that our method achieves state-of-the-art performance without introducing much computational overhead. Code is available at https://github.com/MKSAQW/CPSR.

1. Introduction

The rapid advancement of deep learning (Saltori et al., 2022; Cheng et al., 2023; Chen et al., 2024a; Yaganapu and Kang, 2024; Chen et al., 2024b; Zhang et al., 2024; Chen et al., 2023c; Yao et al., 2021; Wang et al., 2023a, 2022a,b; Lu et al., 2023; Yin et al., 2024; Chen et al., 2023a; Cai et al., 2024) has established the groundwork for enhancing the performance of human visual attention mechanisms (HVAM) (Chang and Zhu, 2023; Hassanin et al., 2024; Luo et al., 2023; Miao, 2022), which are dedicated to simulating the human eye to identify regions of interest in an image. Semantic segmentation (Zhou et al., 2022; Yin et al., 2023a; Chen et al., 2024c; Wu et al., 2024; Li et al., 2022b; Wei et al., 2019; Weng et al., 2024; Zhang et al., 2023), a crucial research direction in HVAM, focuses on classifying each pixel within images. Nonetheless, the efficacy of contemporary semantic segmentation models is significantly dependent upon the volume of labeled data, and acquiring pixel-level labels is exceedingly time-consuming and labor-intensive. Consequently, in recent years, semi-supervised semantic segmentation (Hu et al., 2021; Guan et al., 2022; Tu et al., 2022; Hou et al., 2022; Yin et al., 2023b), which utilizes a small portion of labeled data along with a substantial volume of unlabeled data during training, has attracted increasing attention.

Existing semi-supervised semantic segmentation approaches (Jin et al., 2022; Fan et al., 2022; Ke et al., 2022) primarily improve the utilization of unlabeled data through two perspectives: pseudo-labeling (Yang et al., 2022b; Sohn et al., 2020; Duan et al., 2022; Feng et al., 2022; Liu et al., 2022b; Li et al., 2022a; Duan et al., 2023) and consistency regularization (Oliver et al., 2018; Ouali et al., 2020a; Duan et al., 2024). The key design of consistency regularization-based approaches (Ouali et al., 2020b; Liu et al., 2022a; Zhao et al., 2023) is to apply network, feature, and data perturbations to force the decoder to produce consistent prediction results, thereby enhancing the prediction invariance of the network. The pseudo-labeling based approaches (Chen et al., 2021; Lai et al., 2021; Hu et al., 2021) generate labels for unlabeled data by leveraging the knowledge acquired from labeled data, and subsequently retrain the model by incorporating all data with their corresponding labels. However, the above two types of approaches only utilize pixel information whose predicted confidence exceeds a fixed threshold for supervised training, ignoring the uncertain pixel information below the threshold, leading to limited exploitation of unlabeled data.

[*] Corresponding author.
E-mail addresses: JianJYin_Yu@163.com (J. Yin), starys@sina.com (S. Yan), taochen@njust.edu.cn (T. Chen), cs_chenyi@njnu.edu.cn (Y. Chen), yazhou.yao@njust.edu.cn (Y. Yao).

https://doi.org/10.1016/j.cviu.2024.104146
Received 29 May 2024; Received in revised form 8 August 2024; Accepted 3 September 2024
Available online 5 September 2024

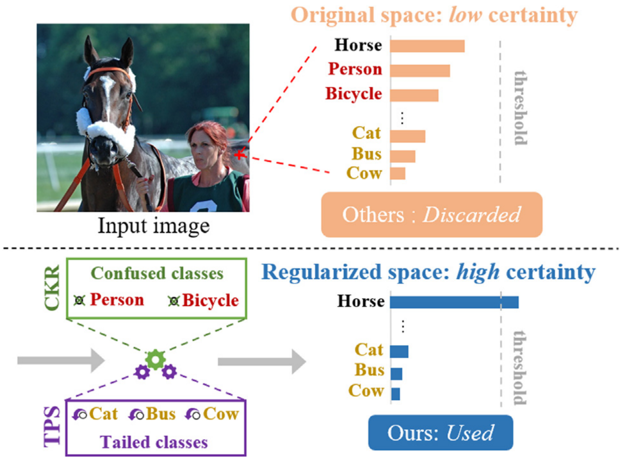

Fig. 1. The motivation of class probability space regularization approach. Unlike other methods that simply discard low certainty pixel information with confidence below a threshold, our approach converts such low certainty data into a usable high certainty one, allowing it to be exploited effectively.

Fig. 1 demonstrates that pixel predictions show uncertainty because the network generates multiple class predictions with similar probability values. The network can determine with high confidence that the ground-truth of the red cross pixel belongs to one of the following classes: *horse*, *person*, or *bicycle*, and the pixel is not predicted as *cat*, *bus*, or *cow*. Essentially, the prediction of the pixel remains ambiguous among several classes with high probabilities, but it is definitively excluded from certain classes.

Based on the above observation, this paper proposes the class probability space regularization (**CPSR**) approach to fully exploit uncertain pixel knowledge for supervising network training. Specifically, a class knowledge reshaping (CKR) module is designed to regularize the original class probability space, thereby converting uncertain pixel predictions into certain ones. We keep the class with the highest probability and gradually remove confused classes, such as the *person* and *bicycle* classes in Fig. 1, until the regularized *Top*-1 probability exceeds a predefined threshold. The real class of the uncertain pixel generally appears in *Top*-1 or confused classes, which means our selected *Top*-1 class may not necessarily be the ground-truth. Therefore, we remove confused classes in the logit space, ensuring that they are not penalized during loss computation and backpropagation. This allows the network to promptly correct classification errors during subsequent training. Intuitively, the class knowledge reshaping module removes confused classes from the original probability space, resulting in a regularized class probability space containing only the *Top*-1 class and tailed classes.

However, when the predicted scores for tailed classes are relatively high, the regularized probability of the *Top*-1 class may still fall below the specified threshold, rendering the pixel unusable. Thus, we further propose the tail probability suppression (TPS) module to counteract the effect of increased tailed class probabilities when constructing the regularized class probability space. Considering the ground-truth is frequently absent from the tailed classes, our proposed TPS applies *hard* cross-entropy loss to the tailed classes in the logit space. Suppressing the tailed class prediction score further promotes the probability value of the *Top*-1 class to surpass the threshold. Our contributions can be summarized as follows: (1) We propose the CPSR method, designed to comprehensively utilize the valuable information from all unlabeled data in network training. (2) The class knowledge reshaping module is proposed to regularize the original class probability space of uncertain pixels and convert them into certain ones for supervised learning. (3) We further propose a tail probability suppression module to handle the problem of ambiguity in the probability for each class in complex scenarios, thereby enabling the model to learn more discriminative information. Extensive experimental results have demonstrated that our approach can fully utilize unlabeled data information, and achieve state-of-the-art performance on the PASCAL VOC2012 and Cityscapes datasets.

The remainder of this paper is organized as follows: Section 2 describes the related work, and Section 3 introduces our approach. In Section 4, we report performance comparisons and ablation studies on two widely used datasets. Finally, we conclude our work in Section 5.

2. Related work

2.1. Semantic segmentation

Semantic segmentation (Guo et al., 2022; Yang et al., 2022a; Kouris et al., 2022; Sun et al., 2020; Lu et al., 2022; Wu et al., 2023a) involves dense pixel-level classification in computer vision. The encoder–decoder architecture, building upon the Fully Convolutional Network (FCN) (Long et al., 2015), has significantly advanced this field. Prior studies (Li et al., 2022b, 2024) have expanded FCN and introduced several innovative methods to capture more comprehensive contextual information. For example, SegNext (Guo et al., 2022) demonstrates that convolutional attention encodes contextual information more effectively than the self-attention mechanism in Transformer. It also designs a novel convolutional attention architecture to enhance the model performance. Mask2Former (Cheng et al., 2022) integrates masking attention into fully convolutional networks, achieving more efficient and accurate image segmentation. GMMSeg (Liang et al., 2022) constructs a Gaussian Mixture Model from a generative perspective to capture class-conditional densities, excelling in both closed-set and open-world scenarios. The LogicSeg (Li et al., 2023) method, based on a holistic visual semantic parser, integrates rich data and symbolic knowledge into the network, enhancing the semantic parsing capabilities of the model. Adaptive masking prompt tuning for masked images is proposed by OVSeg (Liang et al., 2023b) to achieve multi-task weight sharing. GSS (Chen et al., 2023b) introduces a generative learning approach to semantic segmentation, treating it as a problem of generating image-conditioned masks by utilizing latent prior learning processes. Besides, the relationships between pixels are proposed to be modeled by IDRNet (Jin et al., 2024) through the deletion diagnostics module. The unified neural clustering method is designed by CLUSTSEG (Liang et al., 2023c) to handle various image segmentation tasks. PiCo (Zhou and Wang, 2024) proposes a pixel-wise metric learning paradigm to make the embeddings of pixels belonging to the same semantic class more similar. Nevertheless, the described approaches overly rely on a significant volume of labeled data for training the model.

2.2. Semi-supervised semantic segmentation

The objective of semi-supervised semantic segmentation (Wu et al., 2023b; Qin et al., 2022; Yuan et al., 2021; Zhao et al., 2021) (**S4**) is to utilize knowledge gained from labeled data to extract more classification information from unlabeled data. Consistency regularization methods (Yang et al., 2023; Wang et al., 2023b; Chen et al., 2021; Liu et al., 2022a) have become predominant in **S4** research and have demonstrated impressive performance in recent years. CPS (Chen et al., 2021) initializes two identical networks with distinct parameters and compels them to mutually supervise each other, thereby enhancing the utilization of unlabeled data with pseudo-labels. PS-MT (Liu et al., 2022a) introduces an auxiliary teacher branch to create a dual-teacher single-student model. Furthermore, a novel confidence-weighted cross-entropy loss function is designed to enhance the prediction consistency of the model. CCVC (Wang et al., 2023b) seeks to extract valuable insights from areas of conflict between predictions made by two subnetworks, thus enhancing the stability of the model training process. LogicDiag (Liang et al., 2023a) corrects erroneous pseudo-labels by

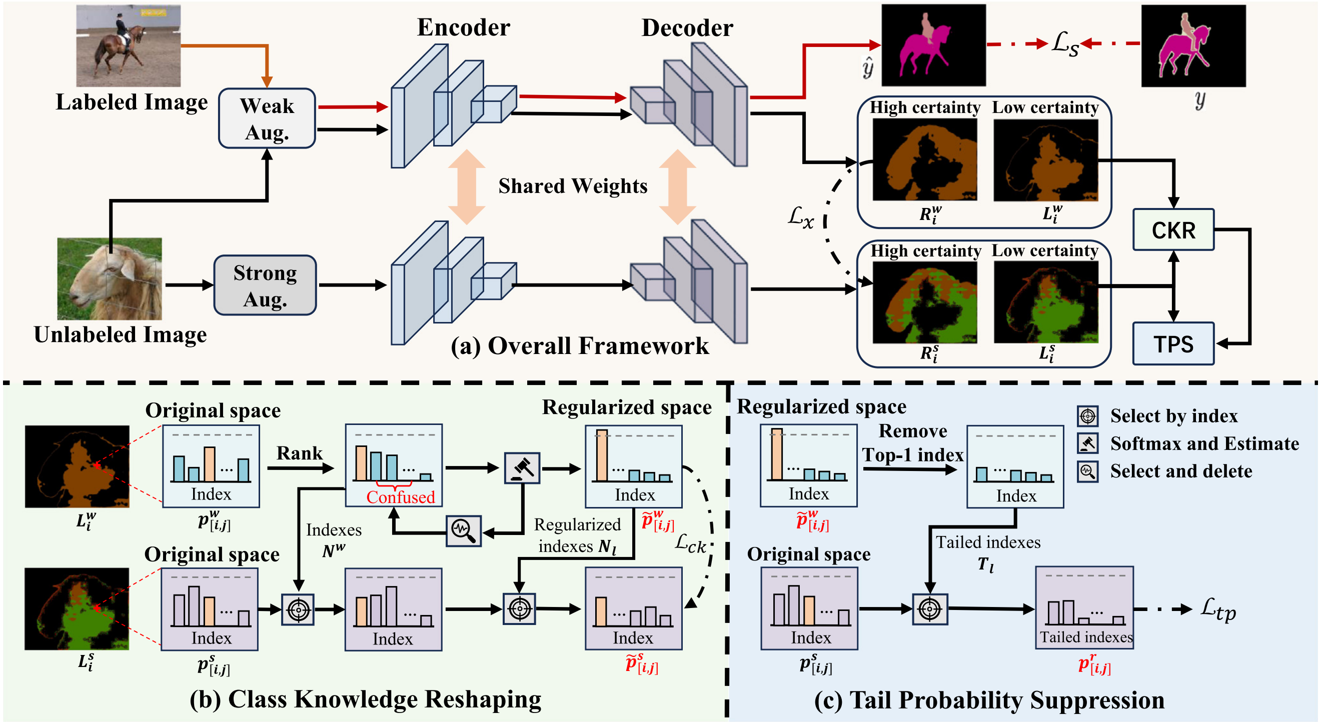

Fig. 2. Overview of class probability space regularization method. The gray dashed line represents the threshold value.

utilizing conflicts identified through symbolic knowledge. Additionally, MKD (Yuan et al., 2023b) designs a framework with two pairs of teacher–student networks and employs cross-supervision for mutual knowledge distillation. UniMatch (Yang et al., 2023) proposes a dual-stream data augmentation strategy, which aims to improve the model's learning of discriminative feature representations, specifically within the perturbation space at the image level. However, the aforementioned methods focus solely on supervising training using pixels from unlabeled data with confidence exceeding a fixed threshold, neglecting the information of uncertain pixels below this threshold. This paper primarily investigates how to effectively utilize uncertain pixel information.

3. The proposed method

In this section, we first introduce the framework structure and training process of the class probability space regularization (CPSR) model. Next, we sequentially introduce the concepts of the class knowledge reshaping (CKR) module and the tail probability suppression (TPS) module, elaborating on how they collaborate to convert low certainty pixels into high certainty ones and participate in supervised training.

3.1. Overall framework

This paper is dedicated to exploring more potential information on massive unlabeled data $S^u = \{(m_i)\}_{i=1}^{|S^u|}$ based on the knowledge learned from limited labeled data $S^l = \{(x_i, y_i)\}_{i=1}^{|S^l|}$. The network structure of our proposed CPSR method is shown in Fig. 2. Our method is also grounded in consistency regularization, with a focus on maintaining prediction consistency. Strong and weak augmentation operations are applied in our approach to generate different augmented versions of the data. This practice improves the model generalization, reduces the risk of overfitting, and enhances the recognition of target region boundaries. Furthermore, our approach concentrates on leveraging the information from low certainty pixels to supervise network training, thereby thoroughly exploring the unlabeled data information. The approach proposed in this paper is illustrated in Algorithm 1.

We initially apply weak augmentation operation to the ith labeled image x_i to generate the corresponding weakly augmented image x^w. Next, we employ ground truth y to supervise prediction generated from weakly augmented image x_i^w, leading to the labeled supervision loss $\mathcal{L}_s$. This loss guides the learning process of the network on labeled data:

$$\mathcal{L}_s = \frac{1}{B_l} \cdot \frac{1}{Z} \sum_{i=1}^{B_l} \sum_{j=1}^{Z} \ell_{ce}(g_t(f_t(x^w_{[i,j]})), y_{[i,j]}). \tag{1}$$

Here, B_l denotes the batch size of labeled images and $\ell_{ce}()$ signifies the cross-entropy loss function. f_t and g_t represent the encoder and decoder, respectively. $Z = H \times W$, where H and W denote the size of the image. $x^w_{[i,j]}$ indicates the jth pixel in the ith image x_i^w. For unlabeled data, we perform strong and weak augmentation operations separately to obtain the corresponding strongly and weakly augmented views (m_i^s and m_i^w). Subsequently, we input both strongly and weakly augmented views into the network to obtain the corresponding predictions (p_i^s and p_i^w). We select high certainty regions R_i^w and low certainty regions L_i^w from the prediction p_i^w of the ith weakly augmented view based on the following formula:

$$\begin{aligned} R_i^w &= \{I(\varphi(p^w_{[i,j]}) > \tau) \cdot p^w_{[i,j]}\}_{j=1}^{Z}, \\ L_i^w &= \{I(\varphi(p^w_{[i,j]}) < \tau) \cdot p^w_{[i,j]}\}_{j=1}^{Z}. \end{aligned} \tag{2}$$

The regions of high certainty R_i^s corresponding to R_i^w and low certainty L_i^s corresponding to L_i^w are identified within p_i^w. $\varphi()$ represents taking the maximum probability value after performing the softmax operation. We input the low certainty regions from the prediction results of weakly augmented views into the class knowledge reshaping module, converting them into high certainty ones. The newly generated high certainty regions are then used to supervise the low certainty regions in strongly augmented views, resulting in the class knowledge reshaping loss $\mathcal{L}_{ck}$. Subsequently, we employ the tail probability suppression module to further enhance the confidence of the *Top*-1 class by suppressing the probabilities of tailed classes, leading to the tail probability suppression loss $\mathcal{L}_{tp}$. Following previous methods (Zhao et al., 2023; Yang et al., 2023), we also use the high certainty regions from the prediction results of weakly augmented views to guide the prediction of strongly

Algorithm 1 CPSR algorithm in a mini-batch.

Input: Labeled batch $\mathcal{B}_l = \{(x_i, y_i)\}_{i=1}^{B_l}$, unlabeled batch $\mathcal{B}_u = \{m_i\}_{i=1}^{B_u}$ $(|\mathcal{B}_l| = |\mathcal{B}_u|)$, weak/strong augmentation $\mathcal{A}_w()/\mathcal{A}_s()$, encoder f_t, decoder g_t, confused class removal and index return function $\vartheta()$, CKR loss calculation $\Phi()$, TPS loss calculation $\varpi()$, maximum value selection $\varphi()$, probability value selection operation based on index and softmax $\Psi()$.

Parameter: threshold τ, CKR loss weight α, TPS loss weight β.

```
$\mathcal{L}_s = \frac{1}{B_l}\cdot\frac{1}{Z}\sum_{i=1}^{B_l}\sum_{j=1}^{Z}\ell_{ce}(g_t(f_t(x^w_{[i,j]})), y_{[i,j]})$   // calculate the supervised loss.
for $m_i \in \mathcal{B}_u$ do
  $p_i^s = g_t(f_t(\mathcal{A}_s(m_i)))$ ,  $p_i^w = g_t(f_t(\mathcal{A}_w(m_i)))$   // obtain segmentation predictions on different augmented images.
  $\mathcal{L}_x = \mathcal{L}_x + \frac{1}{Z}\sum_{j=1}^{Z} I(\varphi(p^w_{[i,j]}) > \tau)\cdot\ell_{ce}(p^s_{[i,j]}, p^w_{[i,j]})$   // high certainty supervision loss.
  $L_i^w = \xi(p^w_{[i]})$   // $\xi()$ aims to select low certainty pixels.
  for $L^w_{[i,j]} \in L_i^w$ do
    $N_l, T_l = \vartheta(L^w_{[i,j]})$   // Obtain the regularized indexes $N_l$ and the tailed class indexes $T_l$.
    $\tilde{p}^s_{[i,j]} = \Psi(L^s_{[i,j]}, N_l)$ ,  $\tilde{p}^w_{[i,j]} = \Psi(L^w_{[i,j]}, N_l)$ ,  $p^r_{[i,j]} = \{p^s_{[i,j]}(k)\}_{k\in T_l}$   // Obtain the regularized prediction results.
    $\mathcal{L}_{ck} = \mathcal{L}_{ck} + \Phi(\tilde{p}^s_{[i,j]}, \tilde{p}^w_{[i,j]})$   // obtain class knowledge reshaping loss $\mathcal{L}_{ck}$.
    $\mathcal{L}_{tp} = \mathcal{L}_{tp} + \varpi(p^r_{[i,j]})$   // obtain tail probability suppression loss $\mathcal{L}_{tp}$.
  end for
end for
return $\mathcal{L} = \mathcal{L}_s + avg(\mathcal{L}_x + \alpha\cdot\mathcal{L}_{ck} + \beta\cdot\mathcal{L}_{tp})$   // return average loss for gradient backpropagation.
```

augmented views:

$$\mathcal{L}_x = \frac{1}{B_u}\cdot\frac{1}{Z}\sum_{i=1}^{B_u}\sum_{j=1}^{Z} I(\varphi(p^w_{[i,j]}) > \tau)\cdot\ell_{ce}(p^s_{[i,j]}, p^w_{[i,j]}). \tag{3}$$

3.2. Class knowledge reshaping

Threshold-based selection methods (Wang et al., 2023b; Yang et al., 2023) lead to the network discarding many uncertain pixels during each training stage, with their predicted confidence falling below a specified threshold (**refer to** Fig. 4(a)), thereby restricting the model's performance. Therefore, the crucial aspect is how to utilize these uncertain pixels effectively.

Intuitively, a straightforward way to utilize uncertain pixels is to use the class with the highest probability value among uncertain pixels for supervised training, as depicted in the following equation:

$$\mathcal{L}_u = \frac{1}{B_u}\cdot\frac{1}{Z}\sum_{i=1}^{|B_u|}\sum_{j=1}^{Z}\ell_{ce}(p^s_{[i,j]}, p^w_{[i,j]}). \tag{4}$$

However, Eq. (4) may cause incorrect pixel classification in the network and penalize the true labels of these pixels in logit space via *hard* cross-entropy, leading to significant inductive bias. We observe that pixel uncertainty arises from confusion within the network among certain *Top* classes. The labels of these uncertain pixels often align with these *Top* classes, which motivates the development of the Class Knowledge Reshaping (CKR) module.

Fig. 2(b) shows the specific process of the class knowledge reshaping module. At first, we sort the prediction results of the low certainty regions L_i^w in the ith weakly augmented view. This sorting is performed in descending order along the channel dimension. The aim is to obtain the class probability values $Q^w = \{q^w_{n_v}\}_{v=1}^{C}$ and the corresponding class indexes $N^w = \{n_v\}_{v=1}^{C}$ for each pixel. In this scenario, C denotes the number of categories in the dataset. During the training of uncertain pixels, we guarantee the perpetual presence of the *Top*-1 class within the original class probability space, gradually removing subsequent confused classes cyclically to regulate the space. This process continues until the probability value of the *Top*-1 class in the regulated class probability space remains above a predefined threshold τ:

$$\varphi(\{q^w_{n_1}\}\cup\{q^w_{n_v}\}_{v=d-1}^{C}) < \tau \ and \ \varphi(\{q^w_{n_1}\}\cup\{q^w_{n_v}\}_{v=d}^{C}) \geq \tau. \tag{5}$$

Confused classes $\{n_v\}_{v=2}^{d-1}$ and tailed classes $\{n_v\}_{v=d}^{C}$ are distinguished by Eq. (5). We then create a regularized class probability space by incorporating the *Top*-1 class and the tailed classes. Next, we adjust the prediction results of the strongly and weakly augmented views based on the class indexes $N_l = \{n_1\}\cup\{n_v\}_{v=d}^{C}$ using the following equation:

$$\tilde{p}^s_{[i,j]} = \Psi(L^s_{[i,j]}, N_l), \quad \tilde{p}^w_{[i,j]} = \Psi(L^w_{[i,j]}, N_l). \tag{6}$$

$\Psi()$ aims to select the probability value of the corresponding position based on the index and softmax operation. Finally, we calculate the class knowledge reshaping loss $\mathcal{L}_{ck}$ to ensure strong and weak consistency prediction:

$$\mathcal{L}_{ck} = \frac{1}{B_u}\cdot\frac{1}{Z}\sum_{i=1}^{B_u}\sum_{j=1}^{Z} I(\varphi(p^w_{[i,j]}) < \tau)\cdot\ell_{ce}(\tilde{p}^s_{[i,j]}, \tilde{p}^w_{[i,j]}). \tag{7}$$

It is crucial to emphasize that the true label typically belongs to the *Top*-1 class and its associated confused categories. In contrast to Eq. (4), we eliminate the confused classes from the original class probability space. This prevents the class knowledge reshaping loss $\mathcal{L}_{ck}$ from penalizing the true label in logit space in instances of misclassification, allowing the network to promptly rectify errors in subsequent training iterations.

3.3. Tail probability suppression

The model produces ambiguous predictions in complex scenarios, leading to minimal differences in probability values among classes. In such cases, the CKR may remove all but the *Top*-1 class from the original class probability space. Therefore, we further propose that the tail probability suppression module learns more useful knowledge from the tailed classes to improve the prediction confidence. Please refer to Fig. 2(c) for details.

The CKR utilizes the Eq. (5) to compute the class index boundary value d for the pixel in uncertain regions and the set of tailed class indexes $\{n_v\}_{v=d}^{C}$. The tail probability suppression module also follows the idea of consistency regularization, transferring the knowledge learned from weakly augmented views to strongly augmented views. We identify the set of tailed classes $T_l = \{k|W^k_{[i,j]} > 0\}_{k=1}^{C}$ corresponding to the low certainty region of the jth pixel in the ith strongly augmented view that satisfies the conditions:

$$W^k_{[i,j]} = I(\varphi(p^w_{[i,j]} < \tau))\cdot I(Rank(p^s_{[i,j]}(k)) > d). \tag{8}$$

The function $Rank()$ sorts probability values in descending order and selects the corresponding rank. The set of tailed class probabilities $p^r_{[i,j]}$ for that pixel can be obtained by tailed class indexes T_l:

$$p^r_{[i,j]} = \{p^s_{[i,j]}(k)\}_{k\in T_l}. \tag{9}$$

Table 1
Performance comparison with other state-of-the-art methods on the ***original*** PASCAL VOC2012 dataset. All methods utilize ResNet-50 as the encoder and Deeplabv3+ as the decoder.

Methods	Publication	Backbone	Performances (%)				
			1/16 (92)	1/8 (183)	1/4 (366)	1/2 (732)	Full (1464)
Supervised	–	ResNet-50	44.03	52.26	61.65	66.72	72.94
PseudoSeg (Zou et al., 2021)	ICLR 2021	ResNet-50	54.89	61.88	64.85	70.42	71.00
CPS (Chen et al., 2021)	CVPR 2021	ResNet-50	64.07	67.42	71.71	75.88	–
GuidedMix-Net (Tu et al., 2022)	AAAI 2022	ResNet-50	–	73.40	75.50	76.50	–
MKD (Yuan et al., 2023b)	ACMMM 2023	ResNet-50	60.60	66.74	71.01	72.73	78.14
CPCL (Fan et al., 2023)	TIP 2023	ResNet-50	61.88	67.02	72.14	74.25	–
AugSeg (Zhao et al., 2023)	CVPR 2023	ResNet-50	64.22	72.17	76.17	77.40	78.82
UniMatch (Yang et al., 2023)	CVPR 2023	ResNet-50	71.90	72.48	75.96	77.39	78.70
ESL (Ma et al., 2023)	ICCV 2023	ResNet-50	61.74	69.50	72.63	74.69	77.11
Ours	–	ResNet-50	**72.20**	**73.02**	**76.61**	**78.31**	**79.44**

Subsequently, we apply the tail probability suppression loss $\mathcal{L}_{tp}$ to reduce the probabilities of the tailed classes, leading to increased probability for the *Top*-1 class and improved prediction accuracy:

$$\mathcal{L}_{tp} = -\frac{1}{B_u} \cdot \frac{1}{Z} \sum_{i=1}^{B_u} \sum_{j=1}^{Z} \sum_{k=1}^{C} W^{k}_{[i,j]} \cdot log(1 - p^{r}_{[i,j]}(k)). \tag{10}$$

3.4. Loss function

We leverage both supervised loss $\mathcal{L}_s$ (defined in Eq. (1)) for labeled data and consistency loss for unlabeled data to train the model. The consistency loss consists of high certainty supervisory loss $\mathcal{L}_x$, class knowledge reshaping loss $\mathcal{L}_{ck}$ and tail probability suppression loss $\mathcal{L}_{tp}$:

$$\mathcal{L} = \mathcal{L}_s + \mathcal{L}_x + \alpha \cdot \mathcal{L}_{ck} + \beta \cdot \mathcal{L}_{tp}. \tag{11}$$

α and β denote the weights of the corresponding losses, respectively.

4. Experiment

This section presents comprehensive comparative experiments of our method with other state-of-the-art approaches on two standard semantic segmentation datasets. Furthermore, we perform detailed ablation experiments on the components of our proposed method.

4.1. Datasets and evaluation metric

Datasets. We use two large datasets to evaluate the performance of our method, including PASCAL VOC2012 (Everingham et al., 2010) and Cityscapes (Cordts et al., 2016).

PASCAL VOC2012 (Everingham et al., 2010) is extensively utilized in computer vision tasks, containing 20 foreground classes and 1 background class. It comprises 1464 images for training and 1449 images for validation. Consistent with the experimental settings of state-of-the-art methods, we employ the ***blended*** PASCAL VOC2012 dataset augmented with the SBD (Hariharan et al., 2011) dataset, which contains 10 582 images. Unlike the ***blended*** PASCAL VOC2012, which selects labeled images from all 10 582 images based on the proportion of labeled data, the ***original*** version targets 1464 finely annotated images for labeled data division, with the remaining images combined with 9118 others to constitute unlabeled images. Building upon the previous works (Yuan et al., 2023b; Zhao et al., 2023; Yang et al., 2023), we compare the performance on the ***original*** PASCAL VOC2012 and ***blended*** PASCAL VOC2012 using the 1/16, 1/8, 1/4, 1/2, Full (1464) and 1/16, 1/8, 1/4 protocols, respectively. In particular, it should be noted that the Full setting indicates that 1464 images are used as labeled data and 9118 images are used as unlabeled data.

Cityscapes (Cordts et al., 2016) is an urban landscape dataset with complex contextual information of 19 classes, containing 5000 fine-grained images. Specifically, 2975 images are used for training, 1525 images are used for testing, and 500 images are used for validation. Following the previous works (Fan et al., 2023; Wang et al., 2023b; Yang et al., 2023), we employ protocols at 1/16, 1/8, 1/4, and 1/2 for performance evaluation.

Evaluation metric. For all the experiments in this paper, we use the Mean Intersection over Union (mIoU) for performance comparison. Given a prediction sequence P and the corresponding ground truth GT, the calculation formula for mIoU is as follows:

$$mIoU = \frac{1}{C} \sum_{i=1}^{C} \frac{\Omega((P = i)\ \&\ (GT = i))}{\Omega((P = i)\ \|\ (GT = i))}. \tag{12}$$

$\Omega()$ counts the occurrences that meet the criteria, while C represents the total number of categories.

4.2. Implementation details

Our method employs an encoder–decoder architecture, using either ResNet-50/ResNet-101 (He et al., 2016) as the encoder and Deeplabv3+ (Chen et al., 2018) as the decoder, to assess its effectiveness. We apply both aforementioned encoders to the PASCAL VOC2012 dataset. However, due to limited computational resources, only ResNet-50 is used for the Cityscapes dataset. The code for our proposed method is based on the Pytorch framework, with training and validation conducted on eight NVIDIA RTX 4090 GPUs. For PASCAL VOC2012, the learning rate is set to 0.001 and the model is trained 80 epochs. Following the prior work (Yang et al., 2023), the ***original*** high quality training set employs the image resolution of 321 × 321, while the ***blender*** set is configured to 513 × 513. For Cityscapes, the learning rate is set to 0.005 and the total training epochs of 240, the input image size is 801 × 801. Both datasets are trained using stochastic gradient descent (SGD) optimization with a batch size of 8, meaning that each iteration processes 8 labeled and 8 unlabeled data.

4.3. Comparisons to the state-of-the-arts

Results on *original* PASCAL VOC2012. Table 1 presents the experimental results of our approach compared to other state-of-the-art methods using the ***original*** PASCAL VOC2012 dataset. As can be seen, our method achieves new state-of-the-art performance across all labeled data ratios. The "Supervised" refers to using only labeled data at the specified ratio for network training, without incorporating unlabeled data. Our approach outperforms the ESL (Ma et al., 2023) by significant margins, achieving higher mIoU across various settings: 1/16, 1/8, 1/4, 1/2, and Full. Additionally, our method also surpasses the UniMatch (Yang et al., 2023) with notable improvements, achieving mIoU gains of 0.3%, 0.54%, 0.65%, 0.92%, and 0.74% mIoU at the same respective settings (Zou et al., 2021).

Results on *blended* PASCAL VOC2012. The comparison results of our method with other state-of-the-art methods on the ***blended*** PASCAL VOC2012 dataset are shown in Table 2. Experimental results are

Table 2
Performance comparison with other state-of-the-art methods on the ***blended*** PASCAL VOC2012 dataset.

Methods	Publication	ResNet-50			ResNet-101		
		1/16 (662)	1/8 (1323)	1/4 (2646)	1/16 (662)	1/8 (1323)	1/4 (2646)
Supervised	–	62.40	68.20	72.30	67.50	71.10	74.20
CCT (Ouali et al., 2020b)	CVPR 2020	65.21	70.90	73.40	67.94	73.00	76.17
CPS (Chen et al., 2021)	CVPR 2021	71.98	73.67	74.90	74.48	76.44	77.68
ST++ (Yang et al., 2022b)	CVPR 2022	72.60	74.40	75.40	74.50	76.30	76.60
U2PL (Wang et al., 2022b)	CVPR 2022	72.00	75.10	76.20	74.40	77.60	78.70
PS-MT (Liu et al., 2022a)	CVPR 2022	72.83	75.70	76.43	75.50	78.20	78.70
AugSeg (Zhao et al., 2023)	CVPR 2023	74.66	75.99	77.16	77.01	77.31	78.82
CCVC (Wang et al., 2023b)	CVPR 2023	74.50	76.10	76.40	77.20	78.40	79.00
UniMatch (Yang et al., 2023)	CVPR 2023	75.80	76.90	76.80	78.10	78.40	79.20
ESL (Ma et al., 2023)	ICCV 2023	73.41	75.86	76.80	76.36	78.57	79.02
Ours	–	**76.11**	**77.16**	**77.46**	**78.54**	**78.86**	**79.37**

Table 3
Performance comparison with other state-of-the-art methods on the Cityscapes dataset. All methods utilize ResNet50 as the encoder and Deeplabv3+ as the decoder.

Methods	Publication	Backbone	Performances (%)			
			1/16 (186)	1/8 (372)	1/4 (744)	1/2 (1488)
Supervised	–	ResNet-50	63.30	70.20	73.10	76.60
CCT (Ouali et al., 2020b)	CVPR 2020	ResNet-50	66.35	72.46	75.68	76.78
CPS (Chen et al., 2021)	CVPR 2021	ResNet-50	69.79	74.39	76.85	78.64
U2PL (Wang et al., 2022b)	CVPR 2022	ResNet-50	69.00	73.00	76.30	77.10
PS-MT (Liu et al., 2022a)	CVPR 2022	ResNet-50	–	75.76	76.92	77.64
CPCL (Fan et al., 2023)	TIP 2023	ResNet-50	69.92	74.60	76.98	78.17
CCVC (Wang et al., 2023b)	CVPR 2023	ResNet-50	74.90	76.40	77.30	–
UniMatch (Yang et al., 2023)	CVPR 2023	ResNet-50	75.03	76.77	77.49	78.60
ESL (Ma et al., 2023)	ICCV 2023	ResNet-50	71.07	76.25	77.58	78.92
Ours	–	ResNet-50	**75.49**	**77.25**	**78.01**	**79.11**

obtained using ResNet-50 and ResNet-101 as encoders. These results demonstrate that our method maintains superior performance regardless of whether ResNet-50 or ResNet-101 is employed as the encoder. Our method outperforms CCVC (Wang et al., 2023b) by 1.61%, 1.06%, and 1.06% mIoU at settings of 1/16, 1/8, and 1/4, respectively, using ResNet-50 as the encoder. Experimental results on both the ***original*** and ***blended*** PASCAL VOC2012 datasets demonstrate the state-of-the-art performance achieved by our method.

Results on Cityscapes. Table 3 presents a comparison of the experimental results between our proposed method and other state-of-the-art methods on the Cityscapes dataset. Our proposed CPSR outperforms UniMatch (Yang et al., 2023), achieving superior performance at settings of 1/16, 1/8, 1/4, and 1/2 with margins of 0.46%, 0.48%, 0.52%, and 0.51% mIoU, respectively. The state-of-the-art performance of our method on both the Cityscapes and PASCAL VOC2012 datasets is attributed to its effective utilization of low certainty pixel information for supervising network training, enabling the network to learn more useful classification information.

Comparison of visualization results. Fig. 3 illustrates the visual segmentation comparison of our method with other state-of-the-art approaches (UniMatch Yang et al. (2023) and AugSeg Zhao et al. (2023)) on the PASCAL VOC2012 dataset. The segmentation results in the first and fourth rows demonstrate that, compared to the others, our approach accurately identifies target regions. Additionally, the results in the second row highlight the excellent performance of our method in complex scenarios, attributed to the tail probability suppression module, which allows the network to learn more discriminative information from the regularized logit space. The failure cases in Fig. 3 reveal the limitations of our method. For instance, the incomplete classification of airplane pixels in the fifth row and the misclassification in the sixth row indicate that our method only considers the information in the logit space of individual pixels, without taking into account the contextual information of neighboring pixels. Although there are some failure cases, extensive experiments have shown that our proposed method still achieves state-of-the-art results at a comprehensive level.

4.4. Ablation studies

Quantitative analysis. Fig. 4(a) provides a quantitative analysis of changes in the ratio of low certainty pixels during the training process using UniMatch (Yang et al., 2023) (**recognized as the Baseline in all ablation experiments**), previously considered the most advanced method. The results reveal that despite training under diverse settings, approximately 10% of low certainty pixels remain unutilized. Our method specifically targets this pixel information loss, striving to achieve complete utilization of low certainty pixels. The performance comparison between our proposed CPSR approach and the *hard* cross-entropy method using the original probability space is shown in Fig. 4(b). Experimental results indicate that using *hard* cross-entropy directly in the original space leads to significant inductive bias. Fig. 4(c) shows the comparison of the ratio of erroneous pixels, where the *Top*-1 class is used as the label, between the regularized probability space and the original probability space during the training process. The data in the figure is derived from the ratio of pixels in the current batch where the *Top*-1 class does not match the true label, relative to the total number of pixels. The figure illustrates that the prediction results obtained after regularizing the original class probability space using the CPSR method are more accurate and reliable. Additionally, the above results demonstrate that employing the regularized space does not penalize the true labels and rather corrects errors promptly during subsequent experiments. This aspect precisely highlights the superiority of our approach.

Parameter Analysis. We conduct detailed experiments on the impact of class knowledge reshaping loss weight α and tail probability suppression loss weight β on model performance, with the experimental results shown in Tables 4 and 5. Setting the loss weight value to zero indicates non-utilization of the corresponding module. Table 4 demonstrates that setting α to appropriate values can consistently enhance the model's performance, with the best experimental performance achieved when α is set to 0.05. The experimental results in Table 5 are obtained by incorporating the tail probability suppression module alongside the class knowledge reshaping loss. Notably, when β is set to 0.005, the

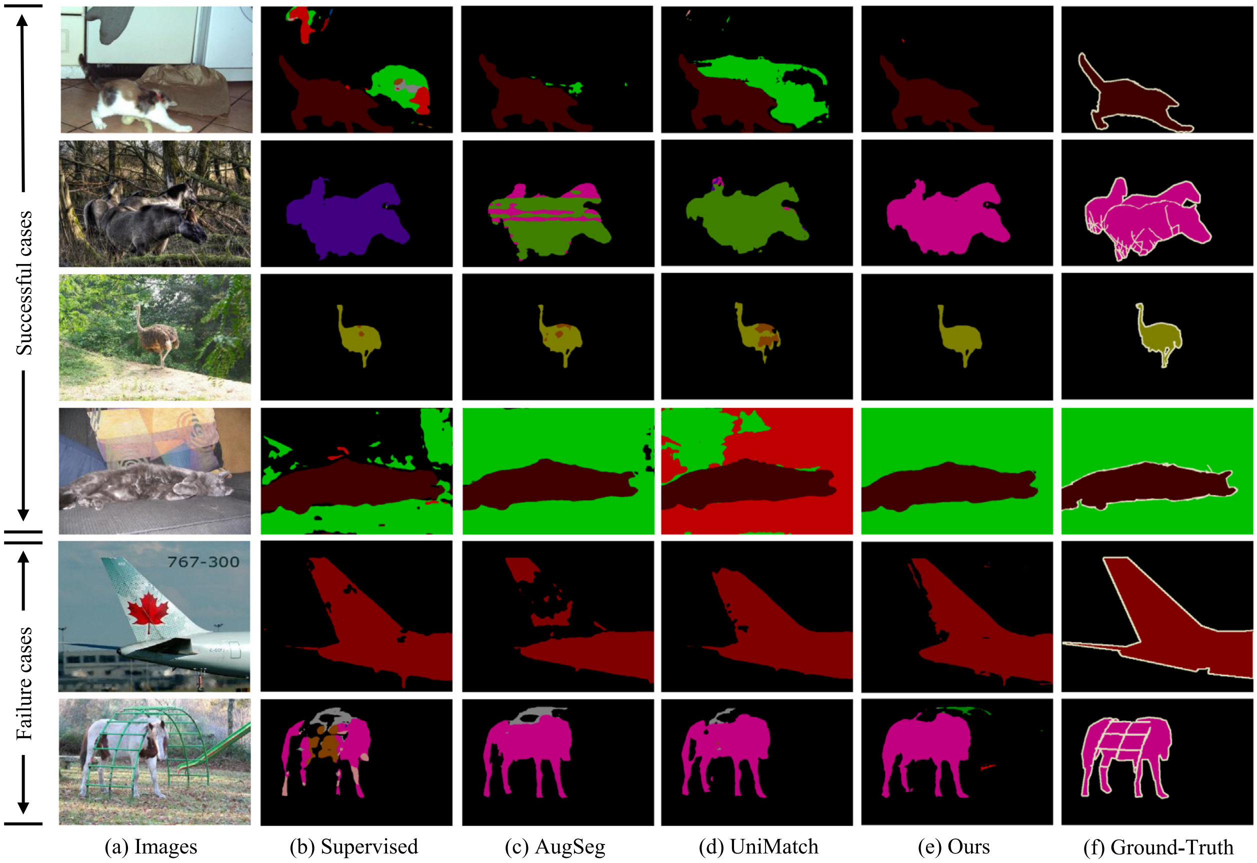

Fig. 3. Comparison of visualization results with other state-of-the-art methods on the PASCAL VOC2012 dataset. All methods are compared using ResNet-50 under the Full (1464) setting for validation. "Supervised" refers to training exclusively with the corresponding proportion of labeled data, without incorporating unlabeled data.

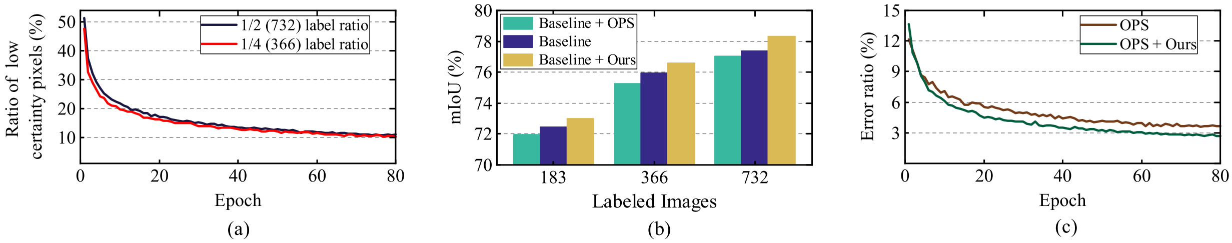

Fig. 4. Quantitative analysis of some ablation experiments on the ***original*** PASCAL VOC2012 dataset. (a) The variation in the ratio of low certainty pixels under different settings of labeled data proportions. (b) Performance comparison between our CPSR method and the *hard* cross-entropy approach using the original probability space (OPS) in Eq. (4). (c) Comparison of the error ratio of the *Top*-1 class used as a label between the regularized probability space and the OPS under 1/4 (2646) setting of the ***blended*** PASCAL VOC2012 dataset.

model achieves a performance of 78.31% mIoU. In other experiments, the values of α and β are set by default to 0.05 and 0.005, respectively.

Additionally, we analyze the threshold τ used for filtering high or low certainty pixels and selecting confused classes, as shown in Table 6. When the value of τ is set to 0, it indicates that the CPSR method is not used. Setting a reasonable threshold significantly enhances the model's performance. For instance, with τ set to 0.9, the performance of the model improves by 0.64% mIoU compared to τ set to 0. The model achieves the best experimental performance at $\tau = 0.95$. It can be observed that removing confused classes in the logit space to transform unreliable pixels into reliable ones, thereby avoiding penalizing true labels, significantly improves the performance of the model. τ is set to 0.95 by default in other experiments.

Impact of each component. Table 7 shows the effect of the class knowledge reshaping (CKR) module and tail probability suppression (TPS) module on model performance. The "Memory" field represents the memory requirement for each GPU when training with two NVIDIA RTX 4090 GPUs under the same settings. The time taken to process a batch of images in each iteration is shown in the "Time" column. The experimental data demonstrate that solely employing the CKR/TPS

Table 4
Impact of class knowledge reshaping loss weight α on model performance under 1/2 (732) setting of the ***original*** PASCAL VOC2012 dataset.

α	0	0.01	0.05	0.1	0.15
mIoU	77.39	77.66	**77.83**	77.71	77.76

Table 5
Impact of tail probability suppression loss weight β on model performance under 1/2 (732) setting of the ***original*** PASCAL VOC2012 dataset.

β	0	0.001	0.003	0.004	0.005	0.006	0.007
mIoU	77.83	78.05	78.11	77.90	**78.31**	78.07	78.29

Table 6
Impact of threshold τ on model performance under 1/4 (2646) setting of the ***blended*** PASCAL VOC2012 dataset.

τ	0	0.8	0.85	0.9	0.95	0.98
mIoU	76.47	77.05	77.04	77.11	**77.46**	77.35

Table 7
Effect of class knowledge reshaping (CKR) module and tail probability suppression (TPS) on model performance under 1/2 (732) setting of the ***original*** PASCAL VOC2012 dataset.

Baseline	CKR	TPS	mIoU	Memory	Time
✓			77.39	15.43G	0.199 s
✓	✓		77.83	15.45G	0.228 s
✓		✓	78.01	15.47G	0.231 s
✓	✓	✓	**78.31**	15.48G	0.236 s

can lead to performance improvements of 0.44%/0.62% mIoU, respectively. We speculate that compared to CKR, TPS offers greater improvement to the model because tailed classes, rather than confused classes, provide more discriminative information. Rational integration of the class knowledge reshaping and tail probability suppression modules enables the model to achieve the best experimental results. Our method prioritizes leveraging pixel information without adding extra parameters, thus not impacting FLOPs. This leads to slight variations in memory usage and image processing time. The data in the table reveals that our method consistently enhances model performance without imposing much computational overhead.

5. Conclusion and limitation

This paper proposes a Class Probability Space Regularization approach to exploit the potential of low certainty pixels in semi-supervised semantic segmentation task. Initially, we design a class knowledge reshaping module that removes confused classes from the logit space, converting low certainty pixels into high certainty ones to supervise the network training. To further learn discriminative information from the logit space, we propose a tail probability suppression module that inhibits the probability distribution of tailed classes, thus significantly augmenting pixel certainty. Extensive experiments and ablation studies were conducted on the PASCAL VOC2012 and Cityscapes datasets to demonstrate the superiority of our proposed approach.

Although our method demonstrates superior performance compared to existing approaches, CPSR solely regularizes the logit space of individual pixels. This conversion of unreliable pixels into reliable ones for network supervision does not account for the contextual information of neighboring pixels. The contextual information of neighboring pixels belonging to the same class within a region can provide rich discriminative information essential for accurate classification. Therefore, our future work will focus on harnessing the contextual information of neighboring pixels for regularization in the pixel class probability space.

CRediT authorship contribution statement

Jianjian Yin: Writing – original draft, Visualization, Validation, Software, Methodology, Formal analysis, Data curation, Conceptualization. **Shuai Yan:** Formal analysis, Funding acquisition, Investigation, Project administration. **Tao Chen:** Supervision, Software, Formal analysis, Data curation. **Yi Chen:** Writing – review & editing, Validation, Supervision, Software. **Yazhou Yao:** Writing – review & editing, Supervision, Software, Resources, Methodology, Investigation, Funding acquisition.

Declaration of competing interest

All authors disclosed no relevant relationships.

Data availability

Data will be made available on request.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 62472222), Natural Science Foundation of Jiangsu Province (No. BK20220934), China Postdoctoral Science Foundation (No. 2022M711635), Jiangsu Funding Program for Excellent Postdoctoral Talent (No. 2022ZB267), Fundamental Research Funds for the Central Universities (No. 30923010303).

References

Cai, X., Lai, Q., Wang, Y., Wang, W., Sun, Z., Yao, Y., 2024. Poly kernel inception network for remote sensing detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 27706–27716.

Chang, Q., Zhu, S., 2023. Human vision attention mechanism-inspired temporal-spatial feature pyramid for video saliency detection. Cogn. Comput. 15 (3), 856–868.

Chen, Y., Huang, W., Liu, X., Deng, S., Chen, Q., Xiong, Z., 2024a. Learning multiscale consistency for self-supervised electron microscopy instance segmentation. In: ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, IEEE, pp. 1566–1570.

Chen, Y., Huang, W., Zhou, S., Chen, Q., Xiong, Z., 2023a. Self-supervised neuron segmentation with multi-agent reinforcement learning. In: Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence. IJCAI-23, pp. 609–617.

Chen, J., Lu, J., Zhu, X., Zhang, L., 2023b. Generative semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7111–7120.

Chen, Y., Shi, H., Liu, X., Shi, T., Zhang, R., Liu, D., Xiong, Z., Wu, F., 2024b. TokenUnify: Scalable autoregressive visual pre-training with mixture token prediction. arXiv preprint arXiv:2405.16847.

Chen, T., Yao, Y., Huang, X., Li, Z., Nie, L., Tang, J., 2024c. Spatial structure constraints for weakly supervised semantic segmentation. IEEE Trans. Image Process. 33, 1136–1148.

Chen, T., Yao, Y., Tang, J., 2023c. Multi-granularity denoising and bidirectional alignment for weakly supervised semantic segmentation. IEEE Trans. Image Process. 32, 2960–2971.

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

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In: European Conference on Computer Vision. pp. 801–818.

Cheng, X., Li, H., Deng, S., Peng, Y., 2023. POEM: A prototype cross and emphasis network for few-shot semantic segmentation. Comput. Vis. Image Underst. 234, 103746.

Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R., 2022. Masked-attention mask transformer for universal image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1290–1299.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B., 2016. The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3213–3223.

Duan, Y., Qi, L., Wang, L., Zhou, L., Shi, Y., 2022. Rda: Reciprocal distribution alignment for robust semi-supervised learning. In: European Conference on Computer Vision. Springer, pp. 533–549.

Duan, Y., Zhao, Z., Qi, L., Wang, L., Zhou, L., Shi, Y., Gao, Y., 2024. MutexMatch: Semi-supervised learning with mutex-based consistency regularization. IEEE Trans. Neural Netw. Learn. Syst. 35 (6), 8441–8455.

Duan, Y., Zhao, Z., Qi, L., Zhou, L., Wang, L., Shi, Y., 2023. Towards semi-supervised learning with non-random missing labels. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 16121–16131.

Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A., 2010. The pascal visual object classes (voc) challenge. Int. J. Comput. Vis. 88, 303–338.

Fan, J., Gao, B., Jin, H., Jiang, L., 2022. Ucc: Uncertainty guided cross-head co-training for semi-supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9947–9956.

Fan, S., Zhu, F., Feng, Z., Lv, Y., Song, M., Wang, F.-Y., 2023. Conservative-progressive collaborative learning for semi-supervised semantic segmentation. IEEE Trans. Image Process. 32, 6183–6194.

Feng, Z., Zhou, Q., Gu, Q., Tan, X., Cheng, G., Lu, X., Shi, J., Ma, L., 2022. Dmt: Dynamic mutual training for semi-supervised learning. Pattern Recognit. 130, 108777.

Guan, D., Huang, J., Xiao, A., Lu, S., 2022. Unbiased subclass regularization for semi-supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9968–9978.

Guo, M.-H., Lu, C.-Z., Hou, Q., Liu, Z., Cheng, M.-M., Hu, S.-M., 2022. Segnext: Rethinking convolutional attention design for semantic segmentation. Adv. Neural Inf. Process. Syst. 35, 1140–1156.

Hariharan, B., Arbeláez, P., Bourdev, L., Maji, S., Malik, J., 2011. Semantic contours from inverse detectors. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. IEEE, pp. 991–998.

Hassanin, M., Anwar, S., Radwan, I., Khan, F.S., Mian, A., 2024. Visual attention methods in deep learning: An in-depth survey. Inf. Fusion 108, 102417.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 770–778.

Hou, J., Ding, X., Deng, J.D., 2022. Semi-supervised semantic segmentation of vessel images using leaking perturbations. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 2625–2634.

Hu, H., Wei, F., Hu, H., Ye, Q., Cui, J., Wang, L., 2021. Semi-supervised semantic segmentation via adaptive equalization learning. Adv. Neural Inf. Process. Syst. 34, 22106–22118.

Jin, Z., Hu, X., Zhu, L., Song, L., Yuan, L., Yu, L., 2024. IDRNet: Intervention-driven relation network for semantic segmentation. Adv. Neural Inf. Process. Syst. 36.

Jin, Y., Wang, J., Lin, D., 2022. Semi-supervised semantic segmentation via gentle teaching assistant. Adv. Neural Inf. Process. Syst. 35, 2803–2816.

Ke, R., Aviles-Rivero, A.I., Pandey, S., Reddy, S., Schönlieb, C.-B., 2022. A three-stage self-training framework for semi-supervised semantic segmentation. IEEE Trans. Image Process. 31, 1805–1815.

Kouris, A., Venieris, S.I., Laskaridis, S., Lane, N., 2022. Multi-exit semantic segmentation networks. In: European Conference on Computer Vision. Springer, pp. 330–349.

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

Li, C., Liu, D., Li, H., Zhang, Z., Lu, G., Chang, X., Cai, W., 2022a. Domain adaptive nuclei instance segmentation and classification via category-aware feature alignment and pseudo-labelling. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 715–724.

Li, L., Wang, W., Yang, Y., 2023. Logicseg: Parsing visual semantics with neural logic learning and reasoning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4122–4133.

Li, L., Wang, W., Zhou, T., Quan, R., Yang, Y., 2024. Semantic hierarchy-aware segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 46 (4), 2123–2138.

Li, L., Zhou, T., Wang, W., Li, J., Yang, Y., 2022b. Deep hierarchical semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1246–1257.

Liang, C., Wang, W., Miao, J., Yang, Y., 2022. Gmmseg: Gaussian mixture based generative semantic segmentation models. Adv. Neural Inf. Process. Syst. 35, 31360–31375.

Liang, C., Wang, W., Miao, J., Yang, Y., 2023a. Logic-induced diagnostic reasoning for semi-supervised semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. ICCV, pp. 16197–16208.

Liang, F., Wu, B., Dai, X., Li, K., Zhao, Y., Zhang, H., Zhang, P., Vajda, P., Marculescu, D., 2023b. Open-vocabulary semantic segmentation with mask-adapted clip. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7061–7070.

Liang, J.C., Zhou, T., Liu, D., Wang, W., 2023c. CLUSTSEG: Clustering for universal segmentation. In: Proceedings of the 40th International Conference on Machine Learning. Vol. 202, PMLR, pp. 20787–20809.

Liu, Y., Tian, Y., Chen, Y., Liu, F., Belagiannis, V., Carneiro, G., 2022a. Perturbed and strict mean teachers for semi-supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4258–4267.

Liu, S., Zhi, S., Johns, E., Davison, A.J., 2022b. Bootstrapping semantic segmentation with regional contrast. In: The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022. OpenReview.net.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3431–3440.

Lu, S., Guo, L., He, X., Zhu, X., Liu, J., Liu, S., 2023. CSDNet: Contrastive similarity distillation network for multi-lingual image-text retrieval. In: International Conference on Image and Graphics. Springer, pp. 385–395.

Lu, L., Xiao, Y., Chang, X., Wang, X., Ren, P., Ren, Z., 2022. Deformable attention-oriented feature pyramid network for semantic segmentation. Knowl.-Based Syst. 254, 109623.

Luo, Y., Lu, Z., Liu, L., Huang, Q., 2023. Deep fusion of human-machine knowledge with attention mechanism for breast cancer diagnosis. Biomed. Signal Process. Control 84, 104784.

Ma, J., Wang, C., Liu, Y., Lin, L., Li, G., 2023. Enhanced soft label for semi-supervised semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1185–1195.

Miao, C., 2022. Research of camouflage evaluation based on human visual attention mechanism. In: International Conference on Computer Research and Development. IEEE, pp. 281–285.

Oliver, A., Odena, A., Raffel, C.A., Cubuk, E.D., Goodfellow, I., 2018. Realistic evaluation of deep semi-supervised learning algorithms. Adv. Neural Inf. Process. Syst. 31, 3239–3250.

Ouali, Y., Hudelot, C., Tami, M., 2020a. An overview of deep semi-supervised learning. arXiv preprint arXiv:2006.05278.

Ouali, Y., Hudelot, C., Tami, M., 2020b. Semi-supervised semantic segmentation with cross-consistency training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12674–12684.

Qin, J., Wu, J., Li, M., Xiao, X., Zheng, M., Wang, X., 2022. Multi-granularity distillation scheme towards lightweight semi-supervised semantic segmentation. In: European Conference on Computer Vision. Springer, pp. 481–498.

Saltori, C., Rota, P., Sebe, N., Almeida, J., 2022. Low-budget label query through domain alignment enforcement. Comput. Vis. Image Underst. 222, 103485.

Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C.A., Cubuk, E.D., Kurakin, A., Li, C.-L., 2020. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. Adv. Neural Inf. Process. Syst. 33, 596–608.

Sun, G., Wang, W., Dai, J., Van Gool, L., 2020. Mining cross-image semantics for weakly supervised semantic segmentation. In: European Conference on Computer Vision. Springer, pp. 347–365.

Tu, P., Huang, Y., Zheng, F., He, Z., Cao, L., Shao, L., 2022. Guidedmix-net: Semi-supervised semantic segmentation by using labeled images as reference. In: Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 36, pp. 2379–2387.

Wang, Y., Huang, S., Gao, Y., Wang, Z., Wang, R., Sheng, K., Zhang, B., Liu, S., 2023a. Transferring clip's knowledge into zero-shot point cloud semantic segmentation. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 3745–3754.

Wang, L., Li, X., Liao, Y., Jiang, Z., Wu, J., Wang, F., Qian, C., Liu, S., 2022a. Head: Hetero-assists distillation for heterogeneous object detectors. In: European Conference on Computer Vision. Springer, pp. 314–331.

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

Wang, Z., Zhao, Z., Xing, X., Xu, D., Kong, X., Zhou, L., 2023b. Conflict-based cross-view consistency for semi-supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19585–19595.

Wei, Z., Zhang, J., Liu, L., Zhu, F., Shen, F., Zhou, Y., Liu, S., Sun, Y., Shao, L., 2019. Building detail-sensitive semantic segmentation networks with polynomial pooling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7115–7123.

Weng, Y., Han, M., He, H., Li, M., Yao, L., Chang, X., Zhuang, B., 2024. Mask propagation for efficient video semantic segmentation. Adv. Neural Inf. Process. Syst. 36.

Wu, W., Dai, T., Huang, X., Ma, F., Xiao, J., 2023a. Top-k pooling with patch contrastive learning for weakly-supervised semantic segmentation. arXiv preprint arXiv:2310.09828.

Wu, W., Dai, T., Huang, X., Ma, F., Xiao, J., 2024. Image augmentation with controlled diffusion for weakly-supervised semantic segmentation. In: ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, IEEE, pp. 6175–6179.

Wu, L., Fang, L., He, X., He, M., Ma, J., Zhong, Z., 2023b. Querying labeled for unlabeled: Cross-image semantic consistency guided semi-supervised semantic segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 45 (7), 8827–8844.

Yaganapu, A., Kang, M., 2024. Multi-layered self-attention mechanism for weakly supervised semantic segmentation. Comput. Vis. Image Underst. 239, 103886.

Yang, L., Qi, L., Feng, L., Zhang, W., Shi, Y., 2023. Revisiting weak-to-strong consistency in semi-supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7236–7246.

Yang, C., Zhou, H., An, Z., Jiang, X., Xu, Y., Zhang, Q., 2022a. Cross-image relational knowledge distillation for semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12319–12328.

Yang, L., Zhuo, W., Qi, L., Shi, Y., Gao, Y., 2022b. St++: Make self-training work better for semi-supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4268–4277.

Yao, Y., Chen, T., Xie, G.-S., Zhang, C., Shen, F., Wu, Q., Tang, Z., Zhang, J., 2021. Non-salient region object mining for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2623–2632.

Yin, J., Chen, Y., Li, C., Zheng, Z., Gu, Y., Zhou, J., 2024. Swin-TransUper: Swin transformer-based UperNet for medical image segmentation. Multimedia Tools Appl. 1–20.

Yin, J., Zheng, Z., Gu, Y., Zhou, J., Chen, Y., 2023a. Class-level multiple distributions representation are necessary for semantic segmentation. arXiv preprint arXiv:2303.08029.

Yin, J., Zheng, Z., Pan, Y., Gu, Y., Chen, Y., 2023b. Semi-supervised semantic segmentation with multi-reliability and multi-level feature augmentation. Expert Syst. Appl. 233, 120973.

Yuan, J., Ge, J., Wang, Z., liu, Y., 2023b. Semi-supervised semantic segmentation with mutual knowledge distillation. In: Proceedings of the 31st ACM International Conference on Multimedia. MM '23, pp. 5436–5444.

Yuan, J., Liu, Y., Shen, C., Wang, Z., Li, H., 2021. A simple baseline for semi-supervised semantic segmentation with strong data augmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8229–8238.

Zhang, S., Ren, G., Zeng, X., Zhang, L., Du, K., Liu, G., Lin, H., 2024. Efficient cross-information fusion decoder for semantic segmentation. Comput. Vis. Image Underst. 240, 103918.

Zhang, L., Zhang, X., Wang, Q., Wu, W., Chang, X., Liu, J., 2023. RPMG-FSS: Robust prior mask guided few-shot semantic segmentation. IEEE Trans. Circuits Syst. Video Technol. 33 (11), 6609–6621.

Zhao, X., Vemulapalli, R., Mansfield, P.A., Gong, B., Green, B., Shapira, L., Wu, Y., 2021. Contrastive learning for label efficient semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10623–10633.

Zhao, Z., Yang, L., Long, S., Pi, J., Zhou, L., Wang, J., 2023. Augmentation matters: A simple-yet-effective approach to semi-supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11350–11359.

Zhou, T., Wang, W., 2024. Cross-image pixel contrasting for semantic segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 46 (8), 5398–5412.

Zhou, T., Wang, W., Konukoglu, E., Van Gool, L., 2022. Rethinking semantic segmentation: A prototype view. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2582–2593.

Zou, Y., Zhang, Z., Zhang, H., Li, C.-L., Bian, X., Huang, J.-B., Pfister, T., 2021. PseudoSeg: Designing pseudo labels for semantic segmentation. In: International Conference on Learning Representations. URL https://openreview.net/forum?id=-TwO99rbVRu.