Multimedia Tools and Applications (2024) 83:8311–8330
https://doi.org/10.1007/s11042-023-15983-w

Weakly supervised semantic segmentation with segments and neighborhood classifiers

Xinlin Xie[1,2] · **Wenjing Zhao**[3] · **Chenyan Luo**[1,2] · **Lei Cui**[1,2]

Received: 18 January 2022 / Revised: 22 May 2023 / Accepted: 4 June 2023 /
Published online: 16 June 2023

Abstract
Semantic segmentation can provide basic semantic information for scene understanding, which has important theoretical research value and broad application prospects. Limited by the labeling cost and the scale of training data, weakly supervised semantic segmentation based on image-level labels has become a potential research issue. However, how to infer the location of image-level labels is a tough problem. Therefore, we propose a weakly-supervised semantic segmentation method with segments and neighborhood classifiers. First, we propose a scheme of segment generation based on the multiple of the number of image-level labels, which can provide high-precision boundary information with fewer regions. Second, to improve the precision of label location inference, we propose an inference method based on the most similar neighborhood granule. It can appropriately determine the number of segments contained in the inferred category label. Finally, we construct a decision table with features as conditional attribute and semantic label as decision attribute, and extract the discriminative features from attribute class reduction for neighborhood classifiers learning. Experiments evidence that our proposed algorithm can produce comparable and competitive results on widely-used MRSC and PASCAL VOC 2012 datasets.

Keywords Semantic segmentation · Image-level labels · Segments · Neighborhood classifiers · Weakly supervised

1 Introduction

Semantic segmentation, i.e., assigning each pixel in the image to one of the pre-defined semantic labels, is a dense pixel-wise prediction task in the field of computer vision [30]. Compared with a single visual task, semantic segmentation can achieve object segmentation and recognition simultaneously. Furthermore, the results of semantic segmentation

✉ Xinlin Xie
xiexinlin@tyust.edu.cn

[1] Taiyuan University of Science and Technology, School of Electronic and Information Engineering, Taiyuan 030024, China

[2] Shanxi Key Laboratory of Advanced Control and Equipment Intelligence, Taiyuan 030024, China

[3] Taiyuan University of Technology, College of Electrical and Power Engineering, Taiyuan 030024, China

can provide fine-grained and high-level semantic information for scene analysis and understanding. Consequently, semantic segmentation can provide core technologies for applications such as automatic driving, intelligent medical, robot perception, 3D urban modeling, intelligent transportation, and remote sensing analysis [38].

Limited by the annotation cost and the scale of training data, weakly supervised semantic segmentation has become a potential research issue. The main forms of weakly supervised annotation include [34]: image-level labels, points, scribbles, and bounding boxes. In particular, weakly supervised semantic segmentation based on image-level labels (i.e., giving only which labels appear in an image, without knowing the location information of each label) is the most popular [26]. The main reasons are: 1) Image-level labels are widely available from media sharing websites, which solves the problem of insufficient training data. 2) Image-level labels can be obtained quickly and efficiently, which greatly reduces the time and cost of data annotation. However, compared with pixel-level labels or other forms of weakly supervised annotation, image-level labels only contain the least guidance information. Specifically, image-level labels do not provide the location information of the label category. Therefore, high-quality and dense label location inference and semantic segmentation model construction based on the inferred pseudo-label data are two key and tough problems. In addition, practical scene image often contains multiple object categories with different appearances and complex backgrounds, which further aggravates the difficulty of label location inference.

According to different ways of label location inference, weakly supervised semantic segmentation based on image-level labels can be divided into proposal-based methods and classification-network-based methods. Proposal-based methods utilize superpixel or segment as processing unit to infer the location of the labels, and then rely on inferred labels for classification model learning. The motivation is that superpixel or segment across multiple images with the same semantic label will have similar appearance [16]. However, the superpixel-level label location inference contains too much redundant information. Compared with superpixels, the number of segments in an image is fewer, which is more helpful for improving the precision of label location inference. However, the existing image segmentation technology cannot ideally segment an image into the number of image-level labels. In addition, classification-network-based methods mainly utilize the pre-trained classification network as prior information to obtain the localization maps of the inferred label, which rely heavily on the precision and universality of pre-trained classification networks. In particular, the classification network only recognizes a small and sparse discriminative region of the object, and cannot provide accurate boundary information of the object.

Although a large number of promising weakly supervised methods have been proposed, the segmentation precision of these methods is still unsatisfactory. The main obstacle is how to accurately and densely implement label location inference (i.e., mapping image-level labels to pixel-level labels) [4]. Therefore, it is necessary to explore a more effective strategy of label location inference. In addition, image-level labels cannot directly give the contour information of the inferred label in the image, which is also an influencing factor for label location inference. Fortunately, segments can be well aligned with the boundaries of an object. Moreover, not all features are equally important and discriminative in the learning phase of the classification model [18].

Motivated by the discussed above, we propose a weakly-supervised semantic segmentation method with segments and neighborhood classifiers (SNC). The overview of our proposed method is shown in Fig. 1. In the part of generating segments, we propose a termination condition based on the multiple of the number of image-level labels for superpixel merging, which can provide high-precision boundary information with fewer regions. In the part of label location inference, we first infer from the label location with

the largest dissimilarity, which guarantees the precision of inference. Second, the number of segments of each inferred semantic label can be appropriately determined based on the multiple and the total number of images in the image set. Then, the neighborhood granule of each segment is constructed, and the segments corresponding to the most similar neighborhood granule constitute the inference of semantic label. In the part of neighborhood classifiers learning, a decision table is constructed by using features of segment as condition attribute and semantic label as decision attribute. Finally, we calculate the significance of each attribute class and learn neighborhood classifiers based on the discriminative features. In the testing phase, we also perform superpixel segmentation and merging on testing images, and use segment as processing unit for prediction. Experimental results on MSRC and PASCAL VOC 2012 datasets demonstrate the effectiveness and comparability of our proposed method in comparison with the state-of-the-arts.

The main contributions of our work are summarized as follows.

- We propose a weakly-supervised semantic segmentation framework with segment as processing unit, which can achieve the segment-level label location inference and prediction.
- We propose a scheme of label location inference based on the most similar neighborhood granule. It can ensure the precision of label location inference and appropriately determine the number of segments in each inferred category label.
- Compared with single attribute, we construct an attribute reduction algorithm with attribute class as feature unit, and use the discriminative features with attribute classes for neighborhood classifiers learning.

2 Related work

2.1 Proposal-based methods

Proposal-based methods use superpixel or segment as processing units for label location inference and prediction. The main implementation steps include: 1) High-precision superpixel segmentation or segment generation; 2) Label location inference with

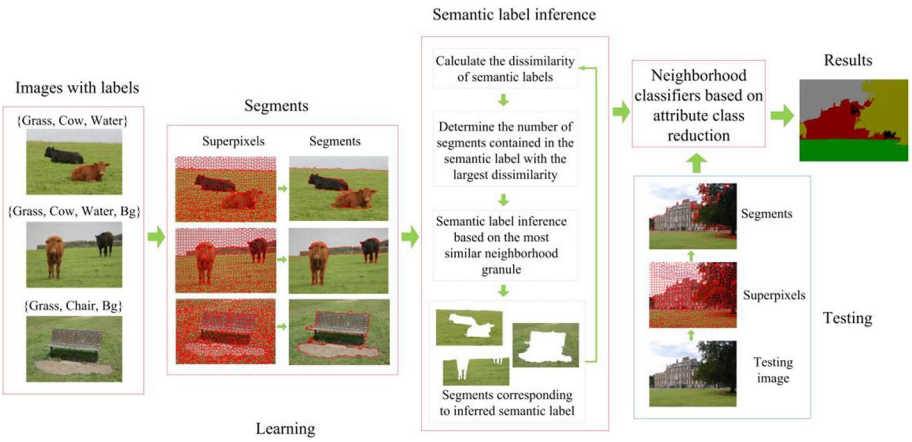

Fig. 1 The overview of our proposed weakly-supervised semantic segmentation framework

superpixel or segment as processing unit; 3) Classification model learning based on inferred pseudo-label data.

The methods using superpixel as processing unit include graph-based methods and clustering-based methods. Among them, graph-based methods mainly use Conditional Random Fields (CRF) or Markov Random Fields (MRF) to construct classification models. For example, Vezhnevets et al. [29] used active learning to find the most informative nodes of CRF; Zhang et al. [41] used Convolutional Neural Network (CNN) to extract multi-scale superpixel features, and used CRF to construct contextual information. Shi et al. [27] constructed the correlation between superpixels based on MRF; Xu et al. [35] used a graph model to encode the presence and absence of categories for superpixel-level label assignment. Clustering-based methods mainly use the feature similarity of superpixels to realize label location inference and prediction. For example, Liu et al. [18] proposed a weakly supervised dual-clustering method by using spectral clustering and discriminative clustering. Liu et al. [19] proposed a multi-instance multi-label learning for dividing superpixels into different clusters. Pourian et al. [23] used spectral clustering to isolate superpixels with high correlation. Zhang et al. [40] used sparse reconstruction to divide the superpixel set, and used iterative merging and updating to obtain the best parameters of the classification model. However, Proposal-based methods rely on the precision of superpixel segmentation and are susceptible to the ability of superpixel feature representation. In addition, image representation and label location inference using superpixel as processing units will result in a large number of redundant superpixels with similar characteristics.

The works most related to ours are [9, 16, 20, 36], which use segment as processing unit. For example, inspired by jigsaw puzzles, Li et al. [16] proposed a semantic segmentation framework based on image piece and CRF. Xu et al. [36] used local search algorithm to merge superpixels into segments, and constructed a unified method based on the clustering framework. For image-level labels with noise, Lu et al. [20] cast weakly supervised semantic segmentation as a noise reduction problem, and used the sparse learning model to detect superpixel noisy labels. Hong et al. [9] proposed a weakly-supervised semantic segmentation algorithm combining CNN and superpixel region response. Compared with superpixels, the methods using segment as processing unit can further reduce the redundant information and computational cost. However, how to adaptively determine the number of segments generated by superpixel merging is a bottleneck problem. A summary of proposal-based methods is shown in Table 1.

2.2 Classification-network-based methods

Classification-network-based methods use pre-trained classification networks to obtain discriminative seed regions. The core links of the methods are: 1) Generating discriminative and sparse seed regions based on the classification network or localization network; 2) Localization and expansion of non-discriminative regions; 3) Segmentation network learning based on pseudo-label data.

Class Activation Map (CAM) localization network proposed by Zhou et al. [42] is widely used in discriminative region mining. On the basis of CAM, Kolesnikov et al. [13] used global weighted pooling to expand the discriminative regions and constructed a boundary constraint model based on fully connected CRF; Araslanov et al. [1] used normalized global weighted pooling to generate dense semantic regions; Saleh et al. [26] used objectness priors with localization network to obtain multi-class masks;

Table 1 Weakly-supervised semantic segmentation methods based on proposals

Classification	Methods	Superpixel segmenttaion	Technology
Superpixels	Vezhnevets et al. [29]	Turbopixels	Active learning, CRF
	Zhang et al. [41]	Multiscale combinatorial grouping (MCG)	CNN, CRF
	Shi et al. [27]	Contour detection and hierarchical segmentation (CDHI)	MRF
	Xu et al. [35] #	CDHI	Graphical model
	Liu et al. [18]	SLIC	Spectral clustering
	Liu et al. [19]	SLIC	Multi-instance multi-label learning
	Pourian et al. [23]	Normalized cut	Spectral clustering, Graph attributed graph
	Zhang et al. [40]	Mean Shift	Sparse reconstruction, Iterative merging update
Segments	Li et al. [16]	SLIC	CRF
	Xu et al. [36]	MCG	Max-margin clustering
	Lu et al. [20]	MCG	Alternate optimization
	Hong et al. [9]	SLIC	CNN, Hierarchical clustering

Code.

Carolina et al. [3] constructed two-class activation graph models to recover the activation mask covering the whole object range; Wei et al. [33] transfer the discriminative information to the non-discriminative object area by setting different dilated convolution rates; Zhou et al. [43] constructed selection loss and attention loss to locate image-level labels and correct classification errors; Wang et al. [30] used paired spatial propagation networks to fine tune the category labels generated by the unit potential network; Lee et al. [14] used FickleNet to identify discriminative regions and non-discriminative regions simultaneously; Huang et al. [11] used seed region growth to expand the discriminative regions. Kho et al. [12] integrated shape and texture information into the mining of discriminative regions, which is an end-to-end semantic segmentation framework. Although CAM-based methods can improve the localization ability of image-level labels, they can only recognize few sparse and incomplete discriminative regions of the image. Moreover, CAM-based methods are prone to generate inaccurate boundary and shape description. Therefore, it is necessary to introduce a subsequent smoothing module to refine the final segmentation result.

Other pre-trained classification models to generate seed regions include attention mechanism, fully convolutional network (FCN), convolutional neural network, saliency detection, and transformer. For example, Wang et al. [31] proposed a self-supervised equivariant attention mechanism to generate dense class activation map information; Li et al. [17] fuse the attention map and saliency map to generate pseudo pixel-level labels. Pathak et al. [21] used FCN and constrained CNN to predict label categories; Qi et al. [24] used FCN to generate the activation map, and combined the object localization network and Multiscale Combinatorial Grouping (MCG) [2] to generate pseudo pixel-level labels. Pinheiro et al. [22] modeled weakly supervised semantic segmentation into a multiple instance learning (MIL) framework and assigned more weight to pixels that are important for classification. Wei et al. [32] use saliency detection to mine saliency maps from

simple images, and use enhanced deep convolutional neural networks (DCNN) to mine pixel-level labels in complex images; Zeng et al. [39] constructed a saliency aggregation module to aggregate the segmentation masks of each prediction category; Fan et al. [6] used instance-level salient object detection to generate candidate regions, and used graph partitioning to construct pseudo-label data. Xu et al. [37] used the transformer model to capture class-specific attention, and utilized discriminative object localization maps as a pseudo-label to achieve weakly supervised semantic segmentation. Ru et al. [25] introduced transformer to get a complete object region, and proposed a weakly supervised semantic segmentation with multi-head self-attention. Although the introduction of prior information can improve the localization precision of the seed region, the segmentation result is sensitive to the universality and versatility of the pre-trained classification network. A summary of classification-network-based methods is shown in Table 2.

3 Proposed method

Based on image-level labels, we propose a weakly-supervised semantic segmentation framework with segments and neighborhood classifiers. As is shown in Fig. 1, our framework consists of two phases: learning and testing. In the learning phase, there are three main parts: 1) Segment generation based on superpixels; 2) Label location inference based on the most similar neighborhood granule; 3) Neighborhood classifiers learning based on discriminative features. In the testing phase, we first perform superpixel segmentation and merging on testing images, and then use segment as processing unit to predict the category label of each pixel.

3.1 Segment generation based on superpixels

In order to obtain segments, we first perform superpixel segmentation on training images, and then merge the superpixels based on low-level visual features until the number of segments is equal to multiple of the number of image-level labels.

In the stage of generating superpixels, we improved the linear spectral clustering (LSC) [15] superpixel segmentation. Specifically, we introduce the detection and reclassification of under-segmented superpixels in the subsequent processing stage of superpixel segmentation. It can not only generate compact and regular superpixels with high precision, but also retain the global property of the image with a linear complexity. The generated superpixels can be represented as $S=\{S(t),\ \ t=1,2,\ldots,K\}$, K is the initial number of superpixels.

In the process of superpixel merging, we first merge the small superpixels to adjacent superpixels based on color and spatial distance. The reason is that small superpixels may have a large contrast with adjacent superpixels, which affects the judgment of termination conditions in superpixel merging. The small superpixels are defined:

$$S(t)\begin{cases} 1, & N(S(t)) \le N/\left(a \bullet t_r\right) \&\& N(S(t)) < \sqrt{N} \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

where, $N(S(t))$ is the number of pixels corresponding to the superpixel; N is the total number of pixels in an image; a is a constant, t_r is the number of superpixels in the merging process.

Table 2 Weakly-supervised semantic segmentation methods based on Classification network

Seed	Methods	Technology	Network structure
CAM	Kolesnikov et al. [13]	DCNN, Fully-connected CRF	VGG-16, DeepLab-CRF-LargeFOV
	Araslanov et al. [1] #	Local consistency, semantic fidelity, completeness, normalised global weighted pooling	VGG-16, ResNet, DeepLabv3+, WideResNet-38 network
	Saleh et al. [26]	Fully-connected CRF	VGG-16, Caffe framework
	Carolina et al. [3] #	Prior network knowledge	VGG-16, DeepLab-CRF-LargeFOV
	Wei et al. [33]	Dilated convolution	VGG-16, DeepLab-CRF-LargeFOV
	Zhou et al. [43] #	Selection and attention losses	VGG-16, ResNet101,
	Wang et al. [30]	Iterative affinity learning, spatial propagation network	DeepLab
	Lee et al. [14]	FickleNet	VGG-16, ResNet DeepLab-VGG16
	Huang et al. [11] #	Seeded region growing	VGG-16, Resnet101, DeepLab-ASPP
	Kho et al. [12]	Texture, CNN, Shape	DeepLabV3, WideResNet-38
Attention	Wang et al. [31]	Self-supervised equivariant attention mechanism	ResNet38, DeepLab
	Li et al. [17]	Attention map, saliency map FH, SLIC, superpixel-CRF	VGG-16, Resnet-101, DeepLab-LargeFOV
FCN	Pathak et al. [21]	Stochastic gradient descent	VGG-16
	Qi et al. [24]	Object localization branch, augmented feedback, MCG	VGG-16, Caffe
CNN	Pinheiro et al. [22] #	Object localization network FH, MCG, MIL	Torch7

Table 2 (continued)

Seed	Methods	Technology	Network structure
Saliency	Wei et al. [32]	Enhanced-DCNN, dense CRF	VGG-16
	Zeng et al. [39]	Saliency aggregation module	VGG-16, DenseNet-169
	Fan et al. [6]	Instance-level salient object detector, graph partitioning	DeepLab, VGG16 ResNet101
Transformer	Xu et al. [37] #	Multi-class tokens, Affinity	DeiT-S Backbone, ResNet38
	Ru et al. [25] #	Multi-head self-attention,	Mix Transformer

Code.

To preserve the boundary information of each superpixel during the merging process, we select low-level visual information to represent the features of each superpixel. Thus, the features of color, shape and texture are selected. Specifically, we choose the mean of LAB and HSV color spaces to represent color features, and use Zernike moments to extract the shape features. Given the ability of modeling frequency and orientation, we chose Gabor filter bank to extract the texture feature of each superpixel.

Based on the similarity between superpixels, the setting of termination condition is the bottleneck of unsupervised image segmentation methods. Fortunately, the number of image-level labels indirectly guides for termination condition. However, there are two deficiencies: 1) Under the existing image segmentation technology, it is difficult to directly generate the same number of segments as the number of image-level labels; 2) Image-level labels only give category label that appears in an image, it will cause under-segmentation when there are non-adjacent objects.

To solve these deficiencies, we propose a termination condition based on the multiple of the number of image-level labels. It solves the problem that the termination condition cannot be adaptively generated during the superpixel merging, and can also generate segments with high precision in practice. Suppose n_L is the number of image-level labels in an image. Therefore, superpixel merging is performed based on the feature similarity between superpixels until the number of superpixels is equal to $k\ n_L$. Where, k is the multiple, which can be adjusted according to the characteristic of training images.

3.2 Label location inference based on the most similar neighborhood granule

Label location inference is the core and key issue in weakly supervised semantic segmentation. With segment as processing unit, we frame the issue as the extraction of the most similar neighborhood granule.

First, the category labels of the dataset can be denoted by $L=[l_1, l_2, \ldots, l_n]$, n is the total number of category labels. From the image-level labels corresponding to training images, we construct an image set $I=\{I(t),\ t=1,2,...,n_i\}$ of each category label, n_i is the number of images. Thus, the relation matrix $R_{m\times n}$ between category labels is defined:

$$R_{m\times n} = \begin{cases} 1, & \text{if the imag e contains the semantic label} \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

where, m is the total number of images in the training set.

Secondly, from image set I and relation matrix R, we can obtain the number of labels for each category label $N_L=\{N_L(t),\ t=1,2,...,n_L\}$ and the second largest number of labels $N_e=\{N_e(t),\ t=1,2,...,n_e\}$. Thus, we define the dissimilarity D of each category label:

$$D(t) = \frac{N_L(t) - N_e(t)}{N_L(T)}, t = 1, 2, \ldots, n \tag{3}$$

From Eq. (3), we encourage the inference from category label that appears in multiple images simultaneously. And then, the inference is started by the category label l_i corresponding to the largest dissimilarity. Furthermore, from the category label l_i and its corresponding image set I_i, we can also get the segment library SL_i corresponding to the category label.

Therefore, the inference of the category label is transformed into finding the segments corresponding to the category label from the segment library. Specifically, the number of segments can be defined:

$$n_s = k' N_L(i), k' \in [1, k] \tag{4}$$

where, k' is a proportional parameter that depends on the multiple k and the complexity of the training images. Essentially, the proportional parameter k' is set to select more similar image sets containing the same label, which will ensure the accuracy of subsequent label location inference. Therefore, we can indirectly obtain the range of the number of segments contained in inferred category label.

To fully represent the features of each segment in the segment library, we adopt the R-CNN [7] and hand-crafted feature extraction manners. R-CNN was proposed and defined by Girshick et al. [7], which denotes regions features with convolutional neural networks. In the R-CNN feature extraction, we extracted the 4096-dimensional feature vector for the circumscribed rectangle of each segment. In the hand-crafted feature extraction, we construct an 81-dimensional feature vector for each segment: LAB color component mean and standard deviation (4-dimensional), HSV color component mean and standard deviation (4-dimensional), shape feature of Zernike moments (7-dimensional), Gabor texture feature mean and standard deviation (2-dimensional), and SURF features (64-dimensional).

Then, based on the feature vector of each segment, we can construct the information table $IS = \langle U, C, V, f \rangle$. U is the universe $\{x_1, x_2, \ldots, x_N, \}$ composed of segment library, N' is the number of segments in the segment library, C is the set of condition attributes (i.e., features), V is the set of attribute values, and f is an information function.

Thus, the neighborhood granule $\delta(x_p)$ of each segment is defined:

$$\delta(x_p) = \{x_q / x_q \in U, \Delta(x_p, x_q) \leq \delta\} \tag{5}$$

$$\Delta(x_p, x_q) = \left(\sum_{j-1}^{N} \left| f(x_1, C) - f(x_2, C) \right|^P \right)^{1/P} \tag{6}$$

where, δ is the neighborhood threshold, which determines the size of the neighborhood granule. Essentially, the neighborhood threshold δ determines the number of segments contained in the label to be inferred. P is the norm. However, when we fixed the size of the neighborhood granule, we can determine which neighborhood granules have the most similar segments. Fortunately, n_S determines the size of the neighborhood granule. Therefore, we can get N' neighborhood thresholds $\delta = \{\delta_1, \delta_2, \ldots, \delta_N, \}$, and then obtain the minimum threshold $\delta_v = \min(\delta)$. Furthermore, the segments within the most similar neighborhood granule $\delta(x_u)$ corresponding to the minimum threshold are determined.

Finally, we obtain the segments corresponding to the category label l_i, and complete the inference of category label l_i. After that, the inferred segments are removed from the segment library, iterated until the inference of all category labels is completed. The complete process for label location inference is outlined in Algorithm 1.

Algorithm 1 Label location inference based on the most similar neighborhood granule

Input: Semantic label set L, image-level labels, segments, k

Output: Labels of segments

1) Construct semantic label set L, image set I, relation matrix R
2) **While** existing uninferred semantic label **do**
3) Calculate the dissimilarity D of each uninferred semantic label by Eq. (3)
4) Obtain the semantic label l_i with the largest dissimilarity, image set I_i, segment library SL_i
5) Determine the segment number n_S contained in the semantic label by Eq. (4)
6) Feature extraction based on R-CNN and hand-crafted features for each segment
7) Construction information table IS
8) Calculate neighborhood granule of each segment by Eq. (5) and Eq. (6)
9) Generate N' neighborhood thresholds $\delta = \{\delta_1, \delta_2, ..., \delta_{N'}\}$ from n_S
10) Obtain the most similar neighborhood granule $\delta(x_u)$ from $\delta_v = \min(\delta)$
11) Complete the label location inference based on the segments corresponding to $\delta(x_u)$
12) Update segment library
13) **end**

3.3 Neighborhood classifiers learning based on discriminative features

How to construct an efficient classification model based on inferred pseudo-label data with segment as processing unit is another important research problem in semantic segmentation. In practice, we also expect to identify pattern with lower dimensional space to avoid the dimensionality disaster. Therefore, we select the neighborhood classifiers [10] to learn the segment sets, which can realize learning based on reduced attributes. However, the attribute reduction of neighborhood classifiers can only deal with a single attribute.

Compared with the single attribute, we propose an improved forward attribute reduction algorithm for neighborhood classifier learning. The proposed method can mine the discriminative features of each segment with the attribute class as the feature unit. The detailed implementation process is as follows.

First, by the segments and their category labels, we build the decision table $DT=\langle U, C\cup D, V, f\rangle$. U is the universe $\{x_1, x_2, \ldots, x_M\}$ composed of inferred segments, M is the number of inferred segments; C is the set of the attributes consisting of the features of segment; D is the decision attribute (i.e., the category of the semantic label). Let attribute class $B \subseteq C$, the significance of B in C is defined:

$$SIG(B, C, D) = \gamma_C(D) - \gamma_{C-B}(D) \tag{7}$$

$$\gamma_B(D) = \frac{\left|POS_B(D)\right|}{|U|} \tag{8}$$

where, $\gamma_C(D)$ is the dependency degree of D to C; $\gamma_{C-B}(D)$ is the dependency degree of the conditional attributes on decision D after dropping attribute B; $POS_B(D)$ is the positive region of decision, which refers to the subset of segments whose neighborhood granules consistently belong to the decision classes; universe U is the set of segments.

Therefore, we can obtain the discriminative features by the improved forward attribute reduction. Finally, the neighborhood classifiers is learned based on the discriminative features and their decision attributes.

Suppose the number of pixels in the image is N and the number of images in the dataset is M. In the segment generation stage, the complexity mainly comes from the superpixel segmentation and superpixel merging. Therefore, the complexity of segment generation is $O(N+(N+(k*n_L)^2))$. In the stage of the label location inference, the complexity mainly includes the classification and inference of category labels. Therefore, the complexity of the label location inference is $O(M*(k*n_L)+L*(2*k*n_L+(k*n_L)^2)$. In the learning phase of the neighborhood classifiers, we use the attribute class as the computational unit. Therefore, the complexity of neighborhood classifiers learning is $O(C*C)$. Thus, the total computational complexity of our proposed algorithm is approximately equal to $O(2N+L*(k*n_L)^2+M*k*n_L+C^2)$.

3.4 Segment-level prediction of testing images

In the testing phase of our method, we still use segment as processing unit for semantic label prediction. The reason is that the segment can adhere well to the boundary of the object and is not susceptible to noise interference.

In order to obtain segments corresponding to the testing images, we first use the same parameter setting and implementation step as the part of segment generation based on superpixels. Secondly, we also perform R-CNN and hand-crafted feature extraction for testing segments, which ensure the consistency between the testing phase and the training phase.

Then, relying on the discriminative features obtained from the attribute class reduction of the training phase, the distance between test segments and semantic labels in the neighborhood classifier is defined:

$$y^* = \min \left\| \varphi^r_{test}(x(i)) - \varphi^r_{train}(x(t)) \right\|, t = 1, 2, \ldots, n \tag{9}$$

where, φ is a metric that represents the label of the segment. Finally, the prediction of the testing segments is output by the semantic labels corresponding to the minimum distance y^*.

4 Experiments

In this section, the proposed weakly-supervised semantic segmentation method based on segments and neighborhood classifiers (SNC) is comprehensively verified and evaluated. First, we describe the datasets used to evaluate the performance of the comparison algorithms. Second, we verify the effectiveness of the proposed superpixel segmentation and segment generation scheme. Finally, the performance of the proposed method and the state-of-the-arts is verified by quantitative and qualitative comparative experiments. We implemented our proposed algorithm on MATLAB R2018a with Intel Xeon Silver 4210 2.10 GHz CPU and 32GB RAM.

4.1 Datasets

MSRC [28]: It is widely used in weakly supervised semantic segmentation, which contained 591 natural scene images from 21 object classes. Each image has an artificially

Table 3 Effect of initial superpixel number K on segmentation performance

	100	200	300	400	500	600	700	800	900	1000
UE	0.21	0.18	0.16	0.14	0.13	0.11	0.11	0.10	0.09	0.09
BR	0.85	0.91	0.94	0.96	0.98	0.99	0.99	1.00	1.00	1.00
ASA	0.93	0.94	0.94	0.95	0.95	0.97	0.98	0.98	0.99	0.99

labeled groundtruth that can be used for the accurate evaluation of experimental results. To ensure the consistency of the experimental results, we split MSRC into 276 training images and 256 testing images.

PASCAL VOC 2012 [5]: It is the most widely used dataset for weakly supervised semantic segmentation, which contains 21 object classes (20 object classes and 1 background class). PASCAL VOC 2012 consists of three subsets: training (1464 images), validation (1449 images) and testing (1456 images). In our experiment, the augmented training set (10,582 images) proposed by Hariharan et al. [8] is adopted, which provides image-level labels for training.

4.2 Performance of superpixel and segment segmentation

To effectively evaluate the performance of superpixel and segment segmentation, we use three standard metrics in the field of superpixel segmentation [15]: under-segmentation error (UE), boundary recall (BR), and achievable segmentation accuracy (ASA). Among them, UE is a metric of boundary adherence, which penalizes the superpixels that do not overlap with the given groundtruth segmentation; BR evaluates the coincidence rate of groundtruth boundary and segmentation boundary; ASA is defined as the upper bound of the object segmentation precision that can be achieved.

First, we verify the effect of initial superpixel number K on the segmentation performance, which has the greatest influence on the improved superpixel segmentation algorithm. The effect on MSRC dataset is shown in Table 3.

At the stage of generating segments, the multiple k is critical to the precision of segment. Under the initial superpixel number $K=1000$, the effect of multiple k on the precision of segment generation on MSRC dataset is verified, which is shown in Table 4.

In order to more intuitively show the effects of parameters K and k on the performance of superpixels and segments, we plot the trend of each metric under different parameters. It is shown in Fig. 2.

As shown in Table 3 and Fig. 2a, as the initial superpixel number K increases, the performance under the three standard metrics continues to increase. The reason is that the image is segmented into more superpixel blocks, which contain more boundary information. However, when K increases to a certain extent, the influence of K on the boundary information is gradually weakened. Moreover, the larger K, the more redundant information is generated in superpixel segmentation. From Table 4 and Fig. 2b, when k is greater,

Table 4 Effect of multiple k on the precision of segment generation

	k=1	*k=2*	*k=3*	*k=4*	*k=5*	*k=6*
UE	0.37	0.30	0.19	0.13	0.10	0.09
BR	0.55	0.68	0.81	0.85	0.87	0.89
ASA	0.77	0.82	0.88	0.91	0.93	0.93

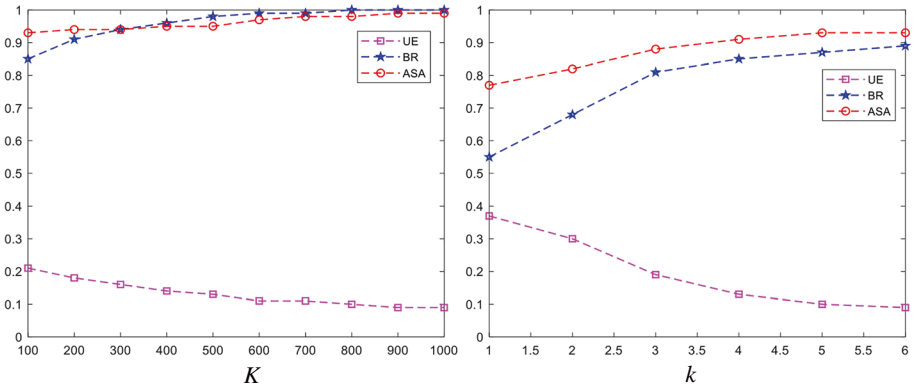

Fig. 2 Effects of parameters K and k on the performance of superpixels and segments

the changes of the three standard metrics become flat. In addition, the three standard metrics increase with increasing k, which proves that increasing k helps to improve the precision of segments. Similarly, the larger the k, the more redundant segments will be generated, which can interfere with the precision of label location inference.

Second, under the condition of $K=1000$ and $k=3$, some segmentation examples of superpixels and segments are shown in Fig. 3.

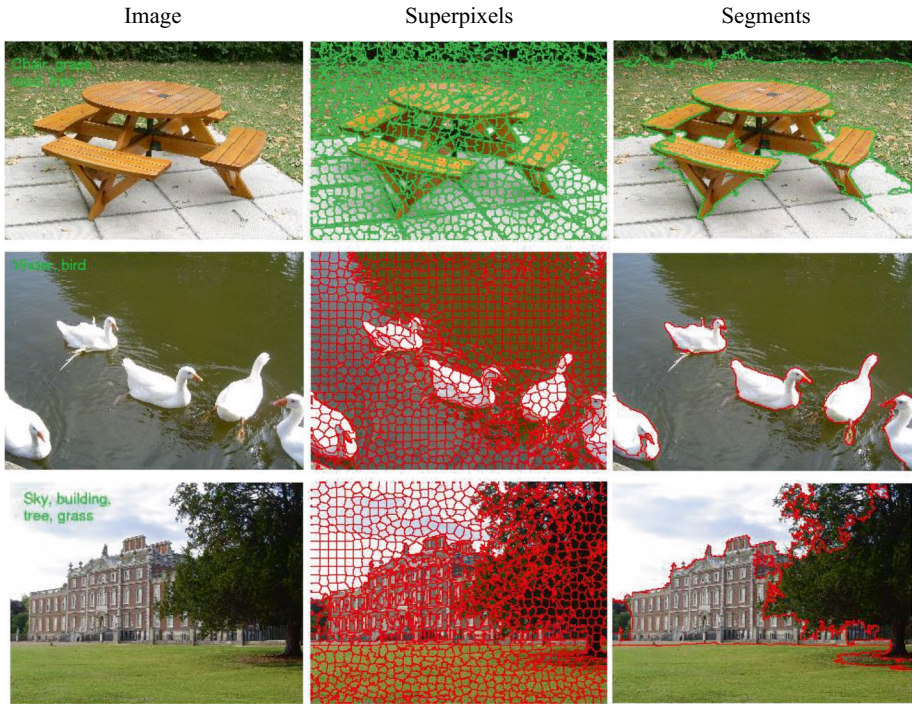

Fig. 3 Some segmentation results when $K=1000$ and $k=3$

As shown in Fig. 3, the proposed segment generation scheme based on multiple of the number of image-level labels can not only effectively avoid the phenomenon of under-segmentation, but also produce fewer segments. Moreover, the proposed termination condition based on the multiple of the number of image-level labels can solve the problem of non-adjacent region annotation with the same label. Furthermore, the generated segments can well preserve the shape and boundary information of the object.

In summary, based on the proposed superpixel and segment generation scheme, we can obtain high-precision superpixels and segments in practice, which helps to improve the precision of label location inference.

4.3 Comparison with state-of-the-art methods

To evaluate the performance of the proposed method (SNC), we compare it with the state-of-the-arts on MSRC and PASCAL VOC 2012 datasets. These comparison methods include: WSDC [18], WNL [20], IPL [16], CCNN [21], SEAM [31], ESC [12], and SAL [43]. All of these methods are weakly supervised semantic segmentation based on image-level labels. In addition, intersection-over-union (IoU) is used to measure the accuracy of segmentation.

First, to quantitatively compare the performance of the proposed algorithm, the per class IoU and mean IoU (mIoU) of SNC and state-of-the art methods on MSRC dataset and the PASCAL VOC 2012 set are shown in Tables 5 and 6, respectively. Bold values show the best performance.

As shown in Tables 5 and 6, the proposed SNC achieves comparable and competitive results on per class IoU and mean IoU compared to state-of-the-arts. Although the total mean IoU is inferior to the IPL [16], SEAM [31], ESC [12], and SAL [43] on MSRC and PASCAL VOC 2012 datasets, the proposed algorithm still achieves the best segmentation performance on some individual labels. However, compared with the supervised information provided by the pre-trained classification networks, the proposed algorithm SNC does not introduce additional priori information. In addition, the segmentation precision of the weakly-supervised semantic segmentation methods on MSRC dataset is significantly higher than the PASCAL VOC 2012 dataset. The reason is that the images in the PASCAL VOC 2012 dataset contain relatively complex objects and background, while the images in the MSRC dataset have more salient objects with larger areas.

Furthermore, to more intuitively visualize the segmentation performance of the proposed method, we give some qualitative segmentation results on PASCAL VOC 2012 dataset, which are shown in Fig. 4.

The segmentation results in Fig. 4 show that our proposed SNC can achieve relatively satisfactory segmentation performance on PASCAL VOC 2012 dataset. Moreover, the segmentation results based on segment-level can preserve the boundary information of an object in the image, which helps to improve the precision of label location inference. However, the proposed superpixel merging scheme will make it difficult to merge segments with large contrast but belonging to the same object. Furthermore, the proposed inference scheme based on the most similar neighborhood granule may also cause wrong inference when the object itself contains multiple segments with high contrast.

Finally, the effect of parameter k' in label location inference on semantic segmentation precision is verified. Because it determines the number of segments contained in each inferred category label, and then affects the segmentation performance of the proposed

Table 5 Precision of per class IoU and mIoU on MSRC dataset (%)

Methods	building	grass	tree	cow	sheep	sky	aeroplane	water	face	car	bicycle	flower	sign	bird	book	chair	road	cat	dog	body	boat	mIoU
WSDC [18]	49	–	58	43	66	–	36	–	46	52	40	85	60	48	–	54	–	52	51	–	–	52.9
MNL [20]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	74.7
IPL [16]	56	92	75	80	88	93	**89**	76	80	**83**	**94**	**99**	**95**	51	**97**	**88**	58	**98**	**83**	58	**86**	81.9
SNC	**78**	**97**	**92**	**86**	**89**	**97**	59	**97**	92	72	57	98	79	**66**	88	80	**91**	48	78	55	58	78.9

- refers to the specific value not given in the original reference.

Bold values show the best performance

Table 6 Precision of per class IoU and mIoU on PASCAL VOC 2012 dataset (%)

Methods	bg	plane	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	motor	person	plant	sheep	sofa	train	tv	mIoU
CCNN [21]	69	26	18	25	20	36	47	47	48	16	38	21	45	35	46	41	30	36	22	39	37	35.3
SEAM [31]	89	69	**33**	**86**	40	**67**	79	**76**	82	29	**76**	**48**	80	74	71	**75**	49	**80**	41	58	53	64.5
ESC [12]	90	76	22	78	**64**	66	80	74	86	30	74	44	**82**	74	64	70	44	79	39	**73**	62	65.2
SAL [43]	**91**	**80**	32	75	54	66	**86**	**76**	**88**	**30**	74	44	80	**79**	**74**	**75**	**50**	79	**47**	58	**63**	66.6
SNC	87	51	8	39	27	44	52	51	34	12	55	17	39	56	49	42	42	60	26	51	44	42.2

Bold values show the best performance

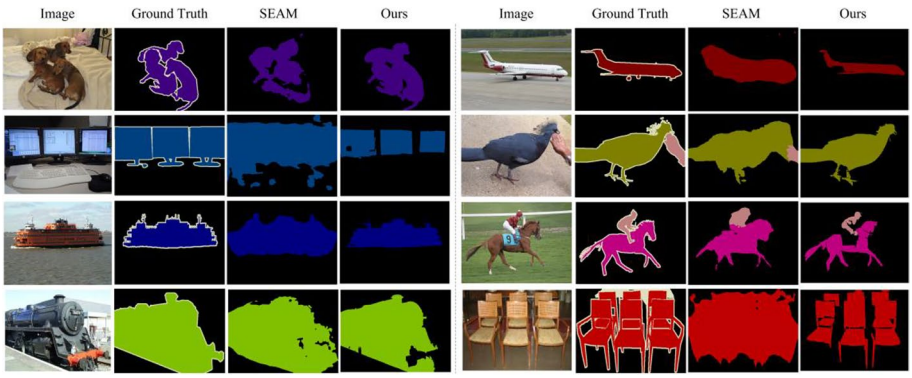

Fig. 4 Some visual segmentation results of the proposed method

method. Therefore, under different parameters *k*', the total mean IoU on MSRC dataset and PASCAL VOC 2012 dataset are shown in Table 7.

As shown in Table 7, under different parameters *k*', the proposed method has relatively large fluctuations in the MSRC and PASCAL VOC 2012 datasets. The main reason is that a smaller *k*' causes fewer segments to be contained in inferred category label, which can affect the learning capability of the classifier. However, a larger parameter *k'* will cause more noisy labels to be selected to the inferred category labels, which will also interfere with the learning of the classifier.

5 Conclusion

We propose a novel framework of weakly supervised semantic segmentation with segments and neighborhood classifiers. The framework uses segment as processing unit to perform label location inference and testing image prediction. In particular, we propose a label location inference scheme based on the most similar neighborhood granule, which can appropriately determine the number of segments contained in inferred category label according to the characteristic of training data. In addition, we implement the extraction of discriminative features based on attribute class reduction of neighborhood classifiers. Experiments on MSRC and PASCAL VOC 2012 datasets show that the proposed method can achieve comparable and promising results compared with the state-of-the-arts. In the future, we will consider introducing some additional prior information into proposal-based methods to improve segmentation accuracy.

Table 7 Effect of parameter k' on segmentation precision (%)

	k'=1	k'=2	*k'=3*	k'=4	k'=5	k'=6
MSRC	62.5	74.3	78.9	78.7	78.5	78.5
PASCAL VOC 2012	25.7	34.3	37.5	41.8	42.2	41.8

Funding None of the authors have any financial or scientific conflicts of interest with regard to the research described in this manuscript.

References

1. Araslanov N, Roth S (2020) Single-stage semantic segmentation from image labels. Proceedings of the IEEE conference on computer vision and pattern recognition. pp 4253–4262
2. Arbeláez P, Pont-Tuset J, Barron JT et al (2014) Multiscale combinatorial grouping. Proceedings of the IEEE conference on computer vision and pattern recognition. pp 328–335
3. Carolina RC, Baptista-Ríos M, López-Sastre RJ (2019) Learning to exploit the prior network knowledge for weakly supervised semantic segmentation. IEEE Trans Image Process 28(7):3649–3661
4. Chang Y, Wang Q, Hung W et al (2020) Weakly-supervised semantic segmentation via sub-category exploration. Proc IEEE conf comput vis pattern recognit. pp 8991–9000
5. Everingham M, Eslami SMA, Van Gool L et al (2015) The pascal visual object classes challenge: a retrospective. Int J Comput Vis 111:98–136
6. Fan R, Hou Q, Cheng M et al (2018) Associating inter-image salient instances for weakly supervised semantic segmentation. Proceedings of the IEEE conference on European conference on computer vision. pp 367–383
7. Girshick R, Donahue J, Darrell T et al (2014) Rich feature hierarchies for accurate object detection and semantic segmentation. Proceedings of the IEEE conference on computer vision and pattern recognition. pp 580–587
8. Hariharan B, Arbeláez P, Bourdev L et al (2011) Semantic contours from inverse detectors. Proceedings of the IEEE international conference on computer vision. 991–998
9. Hong Y, Zhang G, Wei B et al (2023) Weakly supervised semantic segmentation for skin cancer via CNN superpixel region response. Multimed Tools Appl 82:6829–6847
10. Hu Q, Yu D, Xie Z (2008) Neighborhood classifiers. Expert Syst Appl 34(2):866–876
11. Huang Z, Wang X, Wang J et al (2018) Weakly-supervised semantic segmentation network with deep seeded region growing. Proceedings of the IEEE conference on computer vision and pattern recognition. pp 7014–7023
12. Kho S, Lee P, Lee W et al (2022) Exploiting shape cues for weakly supervised semantic segmentation. Pattern Recogn 132:108953
13. Kolesnikov A, Lampert CH (2016) Seed, expand and constrain: three principles for weakly-supervised image segmentation. Proceedings of the IEEE conference on European conference on computer vision. pp 695–711
14. Lee J, Kim E, Lee S et al (2019) Ficklenet: weakly and semi-supervised semantic image segmentation using stochastic inference. Proceedings of the IEEE conference on computer vision and pattern recognition. pp 5267–5276
15. Li Z, Chen J (2015) Superpixel segmentation using linear spectral clustering. Proceedings of the IEEE conference on computer vision and pattern recognition. pp 1356–1363
16. Li Y, Guo Y, Kao Y et al (2017) Image piece learning for weakly supervised semantic segmentation. IEEE Trans Syst Man Cybern Syst 47(4):648–659
17. Li Y, Liu Y, Liu G et al (2020) Weakly supervised semantic segmentation by iterative superpixel-CRF refinement with initial clues guiding. Neurocomputing 391:25–41
18. Liu Y, Liu J, Li Z et al (2013) Weakly-supervised dual clustering for image semantic segmentation. Proceedings of the IEEE conference on computer vision and pattern recognition. pp 2075–2082
19. Liu Y, Li Z, Liu J et al (2015) Boosted MIML method for weakly-supervised image semantic segmentation. Multimed Tools Appl 74:543–559
20. Lu Z, Fu Z, Xiang T et al (2017) Learning from weak and noisy labels for semantic segmentation. IEEE Trans Pattern Anal Mach Intell 39(3):486–500
21. Pathak D, Krahenbuhl P, Darrell T (2015) Constrained convolutional neural networks for weakly supervised segmentation. Proceedings of the IEEE international conference on computer vision. pp 1796–1804
22. Pinheiro PO, Collobert R (2015) From image-level to pixel-level labeling with convolutional networks. Proceedings of the IEEE conference on computer vision and pattern recognition. pp 1713–1721
23. Pourian N, Karthikeyan S, Manjunath BS (2015) Weakly supervised graph based semantic segmentation by learning communities of image-parts. Proceedings of the IEEE international conference on computer vision. pp 1359–1367

24. Qi X, Liu Z, Shi J et al (2016) Augmented feedback in semantic segmentation under image level supervision. Proceedings of the IEEE conference on European conference on computer vision. pp 90–105
25. Ru L, Zhan Y, Yu B et al (2022) Learning affinity from attention: end-to-end weakly-supervised semantic segmentation with transformers. Proceedings of the IEEE conference on computer vision and pattern recognition. pp 16846–16855
26. Saleh FS, Aliakbarian MS, Salzmann M et al (2018) Incorporating network built-in priors in weakly-supervised semantic segmentation. IEEE Trans Pattern Anal Mach Intell 40(6):1382–1396
27. Shi Z, Yang Y, Hospedales TM et al (2016) Weakly-supervised image annotation and segmentation with objects and attributes. IEEE Trans Pattern Anal Mach Intell 39(12):2525–2538
28. Shotton J, Winn J, Rother C, Criminisi A (2006) Textonboost: joint appearance, shape and context modeling for multi-class object recognition and segmentation. Proceedings of the IEEE conference on European conference on computer vision. pp 1–15
29. Vezhnevets A, Buhmann JM, Ferrari V (2012) Active learning for semantic segmentation with expected change. Proceedings of the IEEE conference on computer vision and pattern recognition. pp 3162–3169
30. Wang X, Liu S, Ma H et al (2020) Weakly-supervised semantic segmentation by iterative affinity learning. Int J Comput Vis 128:1736–1749
31. Wang Y, Zhang J, Kan M et al (2020) Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. Proceedings of the IEEE conference on computer vision and pattern recognition. pp 12275–12284
32. Wei Y, Liang X, Chen Y et al (2017) Stc: a simple to complex framework for weakly-supervised semantic segmentation. IEEE Trans Pattern Anal Mach Intell 39(11):2314–2320
33. Wei Y, Xiao H, Shi H et al (2018) Revisiting dilated convolution: a simple approach for weakly-and semi-supervised semantic segmentation. Proc IEEE conf comput vis pattern recognit. pp 7268–7277
34. Wu T, Huang J, Gao G et al (2021) Embedded discriminative attention mechanism for weakly supervised semantic segmentation. Proceedings of the IEEE conference on computer vision and pattern recognition. pp 16765–16774
35. Xu J, Schwing AG, Urtasun R (2014) Tell me what you see and i will show you where it is. Proceedings of the IEEE conference on computer vision and pattern recognition. pp 3190–3197
36. Xu J, Schwing AG, Urtasun R (2015) Learning to segment under various forms of weak supervision. Proceedings of the IEEE conference on computer vision and pattern recognition. pp 3781–3790
37. Xu L, Ouyang W, Bennamoun M et al (2022) Multi-class token transformer for weakly supervised semantic segmentation. Proceedings of the IEEE conference on computer vision and pattern recognition. pp 4310–4319
38. Yao Y, Chen T, Xie G et al (2021) Non-salient region object mining for weakly supervised semantic segmentation. Proceedings of the IEEE conference on computer vision and pattern recognition. pp 2623–2632
39. Zeng Y, Zhuge Y, Lu H (2019) Joint learning of saliency detection and weakly-supervised semantic segmentation. Proceedings of the IEEE international conference on computer vision. pp 7223–7233
40. Zhang K, Zhang W, Zheng Y, Xue X (2013) Sparse reconstruction for weakly supervised semantic segmentation. Proceedings of the international joint conference on artificial intelligence. pp 1889–1895
41. Zhang W, Zeng S, Wang D et al (2015) Weakly supervised semantic segmentation for social images. Proceedings of the IEEE conference on computer vision and pattern recognition. pp 2718–2726
42. Zhou B, Khosla A, Lapedriza A et al (2016) Learning deep features for discriminative localization. Proc IEEE conf comput vis pattern recognit. pp 2921–2929
43. Zhou L, Gong C, Liu Z et al (2020) SAL: selection and attention losses for weakly supervised semantic segmentation. IEEE Transactions on Multimedia 23:1035–1048