Article

Weakly Supervised Conditional Random Fields Model for Semantic Segmentation with Image Patches

Xinying Xu [1], Yujing Xue [1], Xiaoxia Han [1], Zhe Zhang [1], Jue Xie [2,*] and Jinchang Ren [1,3,*]

[1] College of Electrical and Power Engineering, Taiyuan University of Technology, Taiyuan 030024, China; xuxinying@tyut.edu.cn (X.X.); sarah370484224@163.com (Y.X.); hanxiaoxia@tyut.edu.cn (X.H.); zhangzhe@tyut.edu.cn (Z.Z.)

[2] College of Information and Computer, Taiyuan University of Technology, Jinzhong 030600, China

[3] University of Strathclyde, Department of Electronic and Electrical Engineering, Glasgow G4 0LN, UK

* Correspondence: xiejun@tyut.edu.cn (J.X.); jinchang.ren@strath.ac.uk (J.R.)

check for updates

Abstract: Image semantic segmentation (ISS) is used to segment an image into regions with differently labeled semantic category. Most of the existing ISS methods are based on fully supervised learning, which requires pixel-level labeling for training the model. As a result, it is often very time-consuming and labor-intensive, yet still subject to manual errors and subjective inconsistency. To tackle such difficulties, a weakly supervised ISS approach is proposed, in which the challenging problem of label inference from image-level to pixel-level will be particularly addressed, using image patches and conditional random fields (CRF). An improved simple linear iterative cluster (SLIC) algorithm is employed to extract superpixels. for image segmentation. Specifically, it generates various numbers of superpixels according to different images, which can be used to guide the process of image patch extraction based on the image-level labeled information. Based on the extracted image patches, the CRF model is constructed for inferring semantic class labels, which uses the potential energy function to map from the image-level to pixel-level image labels. Finally, patch based CRF (PBCRF) model is used to accomplish the weakly supervised ISS. Experiments conducted on two publicly available benchmark datasets, MSRC and PASCAL VOC 2012, have demonstrated that our proposed algorithm can yield very promising results compared to quite a few state-of-the-art ISS methods, including some deep learning-based models.

Keywords: image semantic segmentation (ISS); weakly supervised; conditional random fields (CRF); image patches

1. Introduction

Different from conventional image segmentation, by combining image segmentation and object recognition, image semantic segmentation (ISS) divides an image into many image blocks to identify the semantic category of each block [1]. It has been widely applied in semantic information extraction from images for scene understanding and object recognition [2,3].

In general, ISS approaches can be mainly divided into two categories, i.e., fully supervised and weakly supervised [4]. Fully supervised ISS requires pixel based labeling of the whole image, which is often achieved manually. To complete the labeling of a picture, skilled annotators on average need nearly 10 min, which is quite time consuming and labor intensive [5]. Considering the difficulty of obtaining pixel-level labeling in fully supervised learning, weakly supervised ISS is more desirable as it does not require pixel based labeling of the whole images thus the associated labor cost and time consumption can be reduced significantly. As a result, weakly supervised ISS has become a research hotspot in recent years.

Although deep learning-based ISS has been recently proposed, such as the Fully Convolutional Networks (FCN) [6], in which an end-to-end pixel-level training is adopted. The FCN uses a 1×1 convolution layer instead of a full connection layer, along with unsampling, for improving the resolution of the feature map. Although FCN has made great success of the segmentation results, it still has some problems. For example, the feature graph is greatly shrunk during the pooling operation, resulting in inaccurate boundary of segmentation [7].

Following the analysis above, in this paper, a novel image patch and conditional random field (CRF) based weakly supervised semantic segmentation (IPCRFWSS) algorithm is proposed. First, we improved the SLIC algorithm to extract the image patches. Second, we construct a second-order CRF with image patches as the node, and the unary and pairwise potential energy functions of CRF are calculated. In addition, the inference of semantic label is transformed into the problem of minimizing potential energy function, and the image patches with semantic labels are obtained. Also, each image patch has been assigned with appropriate semantic labels. Comprehensive experiments on the MSRC datasets have been conducted for quantitative performance evaluation when benchmarking with several state-of-the-art approaches.

The major contributions of our work can be highlighted as follows:

1. We propose an image patch and CRF based weakly supervised ISS algorithm (IPCRFWSS), which can successfully achieve semantic label inference and prediction;
2. We propose an algorithm for automatic estimation of the recommended number of superpixels for different images, which has significantly improved the efficiency and accuracy of image segmentation as it can be used to generate image patches for image-level labels;
3. A PBCRF model is introduced for semantic class inference from image-level to pixel-level labels. With trained patch based CRF, class correlation and similarity functions are added into pairwise potential function to improve the accuracy and robustness of semantic label inference;
4. Experimental results on two publicly available datasets have fully validated the efficacy and efficiency of the proposed approach, which has outperformed quite a few state-of-the-art, including some deep learning models.

2. Related Work

2.1. Conventional Image Segmentation

As a fundamental task for semantic image processing and image understanding, Image segmentation divides an image into different non-overlapped regions according to its color, texture and other visual properties. At present, image segmentation methods can be roughly divided into three categories [8], i.e., region-based [9], edge-based [10], and cluster-based methods [11]. Among them, the region-based segmentation is popularly used.

According to the consistency within the region and the inconsistency between regions, region-based segmentation methods can be further divided into three groups, including thresholding [12], region growing [13], and splitting and merging [14] based techniques. The advantage of thresholding is that it is easy for implementation and the computational complexity is low. However, the spatial position information of the image is ignored, which has led to the difficulty in balancing the segmentation effect in the global and local areas. The region growing method determines a suitable region by using a point as the seed point along with a growing criterion, which is often measured by the similarity of the formed region and the pixel under processing. This method improves the performance of image segmentation, but it is sensitive to noise and can easily lead to over- segmentation [13]. The splitting and merging method divide the image into many small regions by local similarity, where the neighboring small regions can be further merged iteratively if they are sufficiently similar to each other [15].

2.2. Superpixel Based Image Segmentation

As a newly proposed splitting and merging method, superpixel based image segmentation has achieved great progress in recent years. As first proposed by Ren et al. [16], superpixel is defined as a sub-region composed of adjacent pixels with similar texture, color and other visual characteristics. Superpixel based image segmentation is the process of clustering pixels into superpixels, and relevant algorithms can be roughly divided into graph-based and gradient descent-based methods. Graph-based methods mainly include: normalized cut (NC) [17], superpixel lattices (SL) [18], and Felzenszwalb and Huttenlocher (FH) algorithm [19]. Typical gradient descent-based methods are watershed [20], Meanshift [21] and the Simple Linear Iterative Cluster (SLIC) algorithm [22] et al.

Simple linear iterative clustering was first proposed in [22]. Hsu et.al [23] proposed an image segmentation algorithm based on SLIC superpixel, and region merging based on 5-D spectral clustering and boundary-focused region clustering. Ning et.al [24] proposed a novel image segmentation method based on interactive region merging, but users should roughly mark the location and region of the target and background. Gu et.al [25] proposed an algorithm to add the color covariance matrix to the features of superpixel to improve the accuracy of image segmentation.

Compared with pixels, the advantages of superpixels are reflected in two aspects: the calculation is simple, which is helpful to reduce the size of processing objects and the computational complexity of subsequent processing. The number of superpixels can be controlled by adjusting the parameter K, however K needs to be set manually. If K is too large, the advantage of superpixel segmentation will be lost and the unnecessary computational complexity of image segmentation will be increased. If K is too small, the accuracy of image segmentation results will be reduced. Therefore, it is more challenging to manually set the appropriate number of superpixels.

Image patches are merged into larger image regions based on weakly supervised information. Each image patch has only one semantic category, but a target region can be composed of multiple image patches. Compared with superpixel, the proposed method has more advantages in using image patches as the basic unit of weakly supervised ISS. There are two main advantages: the number of image patches is much less than the number of superpixels, which can greatly reduce the complexity of the algorithm. Image patches have more neat object boundaries, which can improve the accuracy of semantic label inference.

2.3. Image Semantic Segmentation

In the past years, ISS has attracted much attention and become one of the hotspots of computer vision. Among the existing ISS methods, there are two main categories: fully supervised and weakly supervised semantic segmentation algorithms. The difference between them is that full supervision requires pixel-level label learning, while weak supervision only needs image-level label learning, which greatly reduces the cost of human and material resources caused by manual annotation [26]. It has a good application prospect, although there are many issues to be addressed in weakly supervised semantic segmentation. These weakly supervised semantics segmentation methods can be roughly divided into two categories: traditional methods and deep learning methods.

Traditional semantic segmentation needs a process of feature extraction followed by several different classifiers to complete the segmentation. Duygulu et al. [27] first proposed the concept of Blob-World, and used image-level label training classifier to conduct image semantic segmentation. Zhang et al. [28] proposed an effective support vector machine classifier based on spatial sparse reconstruction method. The classifier is trained with noisy data and denoised by subspace reconstruction method. The optimal parameters are obtained by iterative optimization. Vezhnevets et al. [29] proposed multi-image model (MIM), using conditional random field model, the one-dimensional potential energy function is established with single superpixel pairs and the two-dimensional potential energy function is established with superpixel pairs. The semantic segmentation result is obtained by CRF parameter approximation solution. Liu et al. [30] proposed weakly-supervised dual clustering for image segmentation and label correspondence. Zhang et al. [31] proposed a graph model for recovering

the pixel based on the appearance similarity of training image superpixels. This model is different from the traditional classifier and has achieved good results when learning multi-class kernel matrices. Wang et al. [32] proposed a probabilistic graph model called TCPR for weakly supervised labeling. This method adds neighborhood context constraint to the MRF model and can use automatic inference mechanism to automatically infer category labels.

Deep learning semantic segmentation method generally consists of a general network framework and a segmentation network. The performance improvement of network structure also brings great improvement to the precision of image processing. To solve the problem of feature loss caused by network framework pooling and down-sampling operation, Noh et al. [33] proposed a deconvolution neural network, which combined the prediction method of network and full convolution network to achieve semantic segmentation task. Farabet et al. [34] proposed multi-scale convolution neural network based deep learning for semantic segmentation, in which pixel-level features were extracted including texture, shape and context information. Qi et al. [35] proposed a framework to reduce the error in weakly supervised learning with image-level supervision. In this way, semantic segmentation and object localization are unified to improve segmentation performance. Wei et al. [36] proposed a framework for generating localization maps by hypotheses-aware classification and cross image contextual refinement. Chen et al. proposed to refine the pixel-wise prediction from the last DCNN layer with a fully connected CRF and achieved better segmentation results [3]. Papandreou et al. [37] develop expectation maximization (EM) methods for semantic image segmentation model training under these weakly supervised and semi-supervised settings. Wei et al. [38] proposed a simple to complex (STC) framework, which used simple image-level labels to enhance the Initial-DCNNs network, and then used the Enhanced-DCNNs network to complete more complex ISS tasks. Although these DCNN-based methods improved the performance of weakly supervised ISS, they rely on the precision of pre-trained classification networks.

3. The Proposed Method

We propose a novel framework for weakly supervised ISS based on image patches and CRF. As shown in Figure 1, the flowchart of the proposed framework contains three main parts, i.e., superpixel generation and image segmentation, CRF model construction, and CRF based semantic inference of image patches for ISS. First, the improved SLIC algorithm is used to segment the training images into superpixels, which are merged into image patches based on the weakly supervised information. Second, the class correlation function and similarity function are introduced into the CRF model to construct a CRF model for inferring semantic class labels. Finally, the trained PBCRF model is applied for weakly supervised semantic segmentation of images. Relevant details of these three parts within the proposed framework are presented as follows.
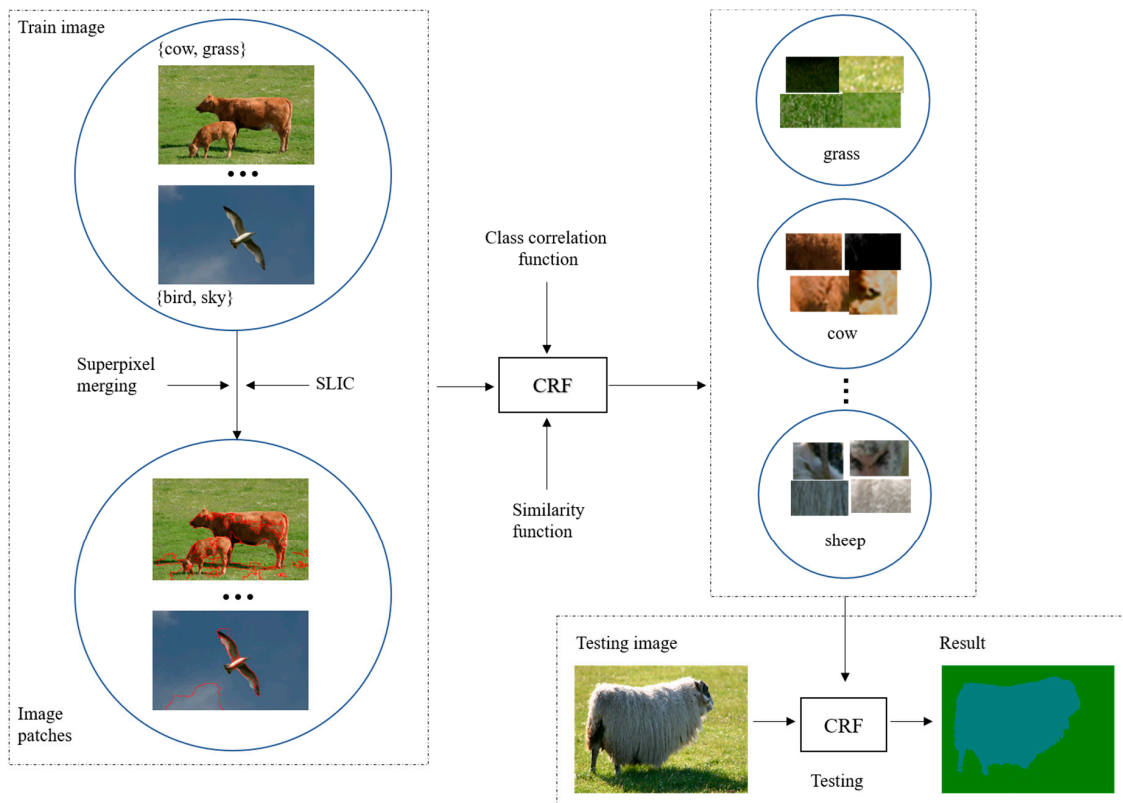
Figure 1. The flowchart of our framework.

3.1. Superpixel Generation

3.1.1. SLIC Superpixel Generation

First, the Simple Linear Iterative Cluster (SLIC) algorithm is used to segment the training images to generate superpixels. Second, superpixels are merged into image patches based on image-level label. The termination condition is that the number of pieces equals three times the number of image-level labels. The better results of superpixel generation and merging are of great help to the construction of CRF model.

Therefore, we have improved the SLIC algorithm so that K can be adaptively determined according to different input images, based on the color information of the images. To better reflect human visual perception, the color space of the image is converted from red, green and blue (RGB) to hue, saturation and value (HSV). In order to simplify the calculation, the H, S, and V components in the HSV space is quantified into 16, 5, and 5 levels respectively, which are further combined to yield a one-dimensional eigenvector Z as follows.

$$Z = HQ_SQ_V + SQ_V + V = 25H + 5S + V \tag{1}$$

where Q_S and Q_V are the quantization grades of S and V respectively ($Q_S = 5$, $Q_V = 5$); the three-dimensional vectors of H, S and V are transformed into one-dimensional vectors ($H \in [0, 360]$, $S \in [0, 1]$, $V \in [0, 1]$). The median value of all elements in Z is determined as m which is used as the initial value for $K' = [m]$. We put m in brackets, as it indicates a rounding up function to ensure an integer value for K'.

3.1.2. SLIC Merging Based on Image-level Labels

In the process, regional feature similarity is taken as the criterion of superpixel merging. Color feature, texture feature and scale invariant feature transform (SIFT) feature are extracted to describe each superpixel. The similarity of superpixels is determined by using the extracted feature vector,

where the adjacent superpixels are merged according to the similarity between different superpixels to obtain the image patches.

Considering the spatial information of two superpixels, i and j, denote $N(i)$ as the neighboring superpixels of i, the adjacency matrix $B(i, j)$ can be defined as:

$$B(i, j) = \begin{cases} 1 \text{ if } j \in N(i) \\ 0 \text{ if } j \notin N(i) \end{cases} \tag{2}$$

In this way, the similarity function can be defined as:

$$\Psi_{i,j} = \begin{cases} \lambda_1 S_{ij}^c + \lambda_2 S_{ij}^t + \lambda_3 S_{ij}^s; \left(\sum_i \lambda_i = 1 \right) & \text{if } B(i, j) = 1 \\ 0 & \text{if } B(i, j) = 0 \end{cases} \tag{3}$$

among them, S_{ij}^c, S_{ij}^t and S_{ij}^s, which are measured Euclidean distances of the color, texture and SIFT feature extracted from two superpixels, i and j, respectively; $\Psi_{i,j}$ denotes the overall distance between the two superpixels, λ_i are the adjusting weights for the three featured.

If $\Psi_{i,j}$ is less than a threshold T, the two superpixels will be merged. The termination condition is set as $P = 3L$, where P is the number of target image patches in the image, and L is the number of labelled categories within the image. The flow chart of superpixel merging algorithm are shown in Figure 2.
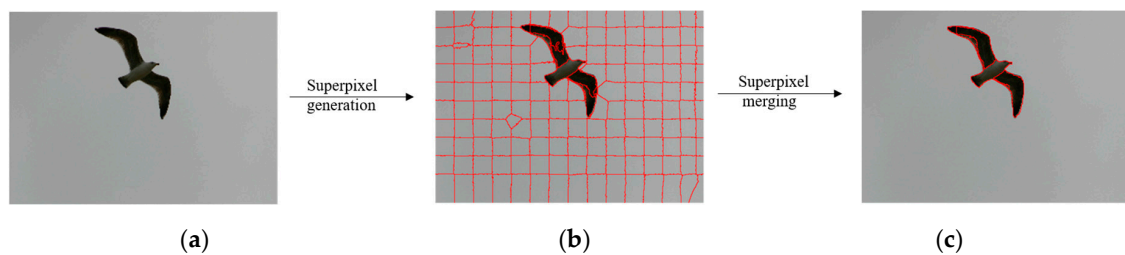

(a) (b) (c)

Figure 2. Illustration diagram of the superpixel merging algorithm: (**a**) the original image, (**b**) the superpixel generated image, (**c**) the superpixel merged into image patches image.

3.2. The Patch Based CRF Model Construct

Semantic inference has always been crucial to weakly supervised ISS as it directly affects the segmentation results. For the determined image patches, they are used as nodes of CRF to construct an undirected graph $G(V, E)$, where V and E denote respectively the node and the edge connecting the nodes. Each image patch needs to be assigned with a unique category of labelled class, which is determined via the CRF model.

On this basis, given the observed field x formed by the extracted patches, the conditional probability distribution of the CRF model for y is defined as:

$$P(y|x) = \propto \exp(-E(x, y)) \tag{4}$$

where $P(y|x)$ is a conditional probability, $E(x, y)$ is an energy function, the final category label assignment result is $\widetilde{y}$ which satisfies the maximum posteriori probability.

$$\widetilde{y} = \text{argmax}_y P(y|x) \tag{5}$$

The energy function of patch based CRF model can be defined as:

$$E(x, y) = w_1 \sum_{i \in v} \phi_i(y_i, x_i) + w_2 \sum_{(i,j) \in E} \phi_{ij}\left(y_i, y_j, x_i, x_j\right) \tag{6}$$

where w_1 and w_2 are weights, $\phi_i(y_i, x_i)$ is a unary potential energy function, which measures the probability that a node i is labeled as y_i for a given x. $\phi_{ij}(y_i, y_j, x_i, x_j)$ is a pairwise potential energy function between adjacent nodes i and j.

Obviously, the solution of y in Equation (5) is the minimum value in Equation (6), thus Equation (5) is equivalent to Equation (7) below.

$$\widetilde{y} = \mathrm{argmin}_y E(x, y) \tag{7}$$

Let $X = [x_1, \cdots, x_P]$ be an image containing P image patches, and x_i is the *i-th* image patch. Its corresponding semantic category is labelled as $y = [y_1, \cdots y_p]$ where $y \in [1, \cdots L]$, and L denotes the total number of categories. However, we can be encoded the label information at the image level, $l(x_i) = [l_1, l_2, \cdots l_L]^T$, where $l_i \in [0, 1]$, and $l_i = 1$ means that the category appears in this image, $l_i = 0$ means not appearing.

The PBCRF model is used to assign similar image patches with the same semantic categories and less similar patches to different semantic categories. In the process of assigning each image patch to an appropriate semantic label, the unary potential energy function of the CRF is formulated as (8)

$$\phi_i(y_i, x_i) = \frac{1}{Z(x)} \exp\left(\sum_{x_i \in N(i)} (1 - l_i(x_i^*, x_i)) \right) \tag{8}$$

where $Z(x)$ is the normalization factor, $N(i)$ refers to the set of image patches adjacent to x_i. l is the value of the image label, and $l_i(\cdot)$ equals to only 0 or 1. $l(x_i^*) \in R^L$ is the true label, x_i is the element of l_i.

Furthermore, in order to assign an appropriate semantic label to each superpixel, category correlation and similarity function are added to the pairwise potential function. Pairwise potential energy functions are defined by

$$\phi_{ij}(y_i, y_j, x_i, x_j) = t(y_i, y_j) \exp\left(-\frac{\|\Psi_{x_i} - \Psi_{x_j}\|^2}{\delta} \right) + (1 - \mu(y_i, y_j)) \exp\left(-\frac{\|D_{x_i} - D_{x_j}\|^2}{\delta} \right) (y_i \neq y_j) \tag{9}$$

where δ is used to adjust the width of the Gauss nucleus, which is set to $\delta = 1$ in the experiment. Ψ_{x_i} is the feature descriptor of x_i, D_{x_i} is the distance feature of x_i.

It is very important to categorize association information for semantic label inference. The category correlation function can be defined by

$$t(y_i, y_j) = P(l(x_i)|l(x_j)) \frac{P(l(x_i)l(x_j))}{P(l(x_j))} \tag{10}$$

Let $l = [l_{x_1}, l_{x_2}, \cdots l_{x_p}]^T \in R^{P \times L}$ be the category label of the image and L the total number of label categories In Equation (10), $P(l(x_i)l(xj))$ is the probability of both class labels $l(x_i)$ and $l(x_j)$, and $P(l(x_j))$ is the probability of the class labels $l(x_j)$.

At the same time, cosine similarity function is used to test the similarity between semantic categories:

$$\mu(y_i, y_j) = \frac{\sum\limits_{i=1} (l(x_i) \cdot l(x_j))}{\sqrt{\sum\limits_{i=1} l(x_i)^2} \times \sqrt{\sum\limits_{i=1} l(x_j)^2}} \tag{11}$$

In this way, the semantic label inference is transformed into the energy function of minimizing conditional random fields, and the semantic category of each image patch is the result of minimizing the energy function.

3.3. CRF Based Semantic Inference of Image Patches for ISS

After each image patch is assigned the appropriate semantic label, the image patches belonging to the same class are put together. According to Equation (6), the PBCRF model is constructed, and the mapping issue between category labels and image patches is transformed into a problem of minimizing the energy function. The main steps of CRF based semantic label inference are shown in **Algorithm 1**:

Algorithm 1: Semantic label inference based on CRF

Input: Image patches P of training images, image-level semantic label and parameters
Output: Semantic Segmentation Results $\hat{y}$
Step 1: Random arrangement of training images
Step 2: Constructing undirected graph $G(V, E)$ with superpixels as nodes
Step 3: Calculating unary potential energy function $\phi_i(y_i, x_i)$ according to Equation (8)
Step 4: The class correlation function $t(y_i, y_j)$ and cosine similarity function $\mu(y_i, y_j)$ are calculated by Equation (10) and Equation (11), and the pairwise potential energy function $\phi_{ij}(y_i, y_j, x_i, x_j)$ is calculated by Equation (9).
Step 5: Constructing potential energy function of patch based on CRF by Equation (6)
Step 6: The semantic segmentation result $\hat{y}$ can be obtained by minimizing the potential energy function

---

In Section 3.2, we add the category correlation and similarity information to the pairwise potential energy function to for semantic label inference.

$$\hat{y} = \text{argmin}_y E(x, y) \tag{12}$$

4. Experiments and Discussion

In this section, comprehensive experiments on two publicly available datasets, MSRC and PASCAL VOC 2012, are used to evaluate the performance of our proposed IPCRFWSS method for ISS. Relevant details including the description of the datasets, parameter settings and benchmarking with several state-of-the-art approaches are presented as follows.

4.1. Dataset Description

Comparative experiments were conducted on two standard datasets, MSRC and PASCAL VOC 2012, both are multi-class data sets including many common natural scenes as detailed below.

MSRC: A multi-class dataset which contains 591 pictures in 21 categories, of which ~80% of the pictures have multiple categories. We divide the dataset into training and test sets according following the same way in [39]. As shown in Figure 3, the 21 categories include: aeroplane, building, bike, bird, book, body, boat, cow, car, chair, cat, dog, face, flower, grass, road, sheep, sky, sign, tree, and water.

Figure 3. A set of images and their corresponding ground-truth annotations.

PASCAL VOC 2012: Serving as the segmentation benchmark for weakly supervised ISS for years [40], this dataset contains 20 object categories and one background category. It contains three parts: training (1464 images), validation (1449 images) and testing (1459 images). In this multi-class

dataset, almost each image has 2 to 4 categories, where most of the images have a complicated background. Figure 4 shows for example some typical images and the corresponding ground-truth. These pictures can be subdivided into four main categories [41] i.e., Vehicles: aeroplane, bicycle, bus, car, motorbike, and train; Animals: bird, cat, cow, dog, horse, and sheep; Household: bottle, chair, dining table, potted plant, sofa, and TV/monitor; and Person, including adults and children though these are not explicitly labelled.

Figure 4. A set of images and their corresponding ground-truth annotations.

4.2. Parameter Settings and Evaluation

To show more intuitively the effect of initial superpixel K on superpixel segmentation, the three most widely used evaluation indicators are adopted. Boundary recall (BR) measures is the proportion of the target boundaries recovered by the superpixel boundaries. Achievable segmentation accuracy (ASA) is a performance upperbound measure. Under-segmentation error (UE) is an error generated by the algorithm when the image is segmented compared with ground truth. The definition of evaluation index used in [42] is adopted here. We compared the number of initial superpixels manually set on MSRC datasets, which visualizes the effect of the number of initial superpixel K on the result of superpixel segmentation. The effect of the initial number of superpixels K is shown in Figure 5.
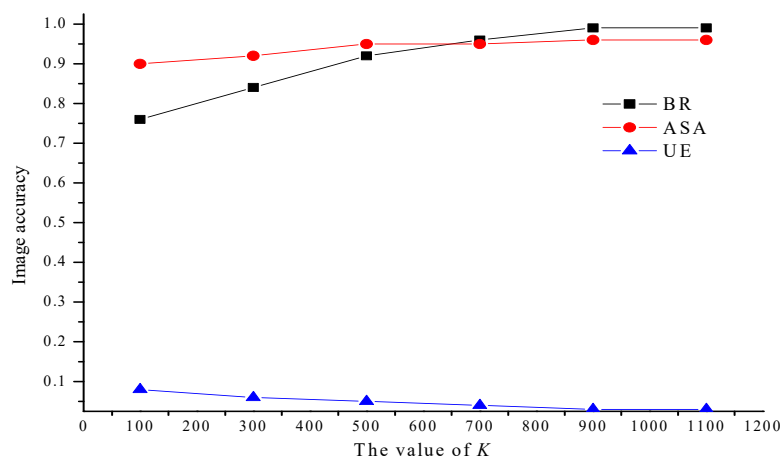

Figure 5. The effect of the initial number of superpixels K on image accuracy.

As shown in Figure 5, from the three evaluation indicators, the performance of image segmentation also improves with the increase of the initial number of superpixels K in a specific range. After that, it will be saturated later. In this state, although the value of K continues to increase, the segmentation performance remains basically unchanged. If the value of K is too large for the image segmentation, it will increase redundant information and the complexity of subsequent calculation.

The accuracy of superpixel segmentation directly affects the results of subsequent ISS, and the value of K determines the size and number of superpixels. If the value of K is too small to achieve good segmentation results, while too large K will bring redundant information.

Therefore, the appropriate *K* helps to achieve good segmentation results. For visual assessment, we conduct a comparison experiment on the number of superpixels *K* in SLIC superpixel segmentation. As shown in Figure 6 and Table 1, the proposed algorithm is feasible in generating satisfactory results on different cases.
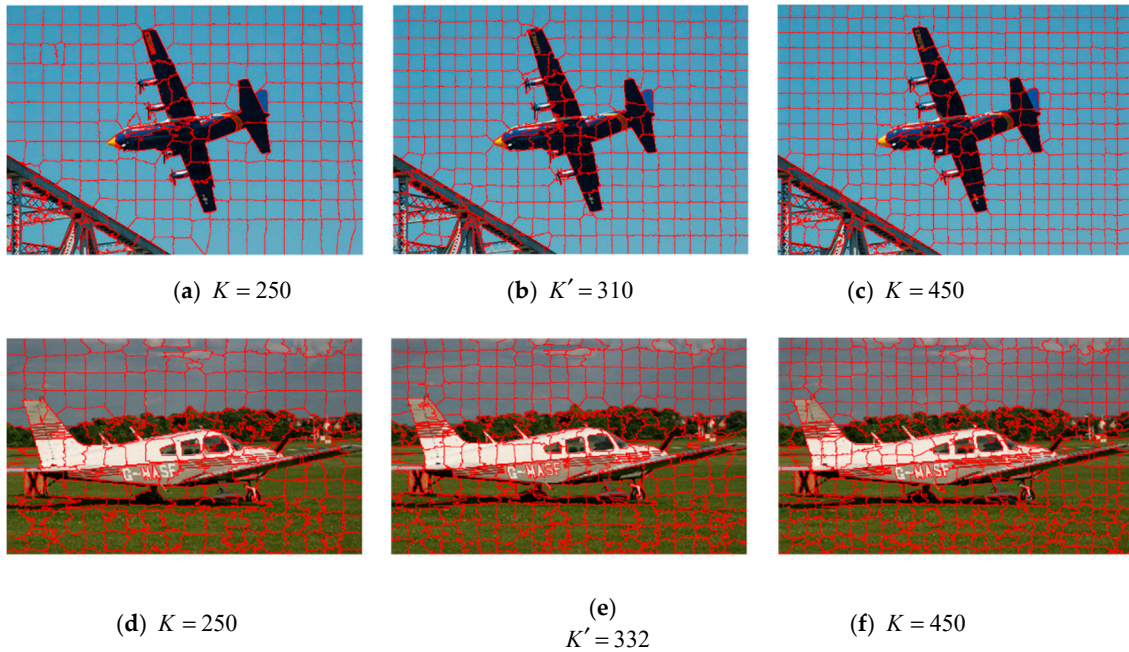

(**a**) *K* = 250   (**b**) *K′* = 310   (**c**) *K* = 450

(**d**) *K* = 250   (**e**) *K′* = 332   (**f**) *K* = 450

Figure 6. Comparison of results under different numbers of initialized superpixels, where *K* and *K′* indicate respectively the number of superpixels set manually and generated by the proposed method.

Table 1. The effect of different *K* values on the precision of segmentation.

Test Image and Size	Manually Set K	F_1	Determined $K^{'}$	F_1
Image 1 (320 × 213)	150	0.829	275	0.832
	350	0.816		
Image 2 (320 × 213)	250	0.834	320	0.857
	450	0.828		
Image 3 (350 × 375)	250	0.823	323	0.845
	450	0.853		
Image 4 (500 × 375)	150	0.852	280	0.841
	450	0.834		

Figure 6b,e show the results of image segmentation from our algorithm with the parameter *K* being set as 310 and 332, respectively. To better compare the effectiveness of the algorithm, the experiment compares the *K* obtained by our proposed algorithm with the segmentation result of *K* around 100 (*K* is set manually with traditional SLIC). Figure 6a is the result of *K* = 250, the head and tail of the plane are not well segmented. Figure 6d shows the Figure 6 the segmentation result of *K* =250. The outline of the plane and some details are not very good. Figure 6b,e have better segmentation effects.

In Table 1, although the number of superpixels is increasing, the F_1-score has not increased significantly. The proposed method can determine the number of superpixels according to different images without multiple attempts to determine appropriate superpixels, which thus saves the running time cost and improves the efficiency.

4.3. Comparing with the State-of-the-Art

Here, we perform a group of experiment to evaluate the performance of weakly supervised ISS method. For reference, we compared the state-of-the-art methods, such as PLSA [43], WSDC [30], Textonboost [39],

MIM [29]. Table 2 shows the segmentation performance on MSRC dataset. The proposed algorithm is compared with CCNN [44], EM-Adapt [37], MIL-ILP-seg [45], SN-B [36] and H&M [46]. The experiments of these methods are carried out on PASCAL VOC 2012 dataset, and the performance of image segmentation is shown in Table 3. The performance is measured in terms of pixel intersection-over-union (IoU) and mean intersection-over-union (mIoU) across 21 classes. In Tables 2 and 3, each column represents the accuracy of each semantic class, and the last column is the average accuracy of all classes. For the values in the table, the values in bold represent the best segmentation performance of this category.

As seen in Table 2, our method provides competitive results when compared with the state-of-the-art methods on the MSRC dataset. Although the accuracy is not as high as others in some categories, yet the overall mIoU is the best among the group. Actually, our approach has produced the best results in six categories, whilst the second and the third overall best approaches, MIM and Textonboost, are dominant in eight and six categories, respectively. This shows our approach can balance in between different classes for good overall performance.

For the results on the PASCAL VOC 2012 dataset in Table 3, our results are the second best in terms of mIoU, which is quite comparable to the best one produced by SN-B, a deep learning-based model. However, our approach significantly outperforms another deep learning model, CCNN, and two other approaches, MIL-ILP-seg and EM-Adapt. Actually, SN-B produces the best results in nine categories of objects, whilst our approach generates the best in seven categories, although the overall mIoU is only 1% lower. Again, the proposed approach seems to be more robust over different categories.

To show the performance of our proposed algorithm more intuitively, extensive experiments were performed on the PASCAL VOC 2012 dataset. As shown in Figure 7, the experimental results are compared with the ground truth. It can be seen from the comparison of segmentation results in Figure 7. that better segmentation results can be achieved when the image object contains only one dominant (merged) superpixel or if the background is relatively simple. On the contrary, when the background of the image is more complex, the accuracy of ISS will also be reduced. In addition, for example, there are many objects in the image, so that occlusion or small shadows between these objects will affect the result of ISS.
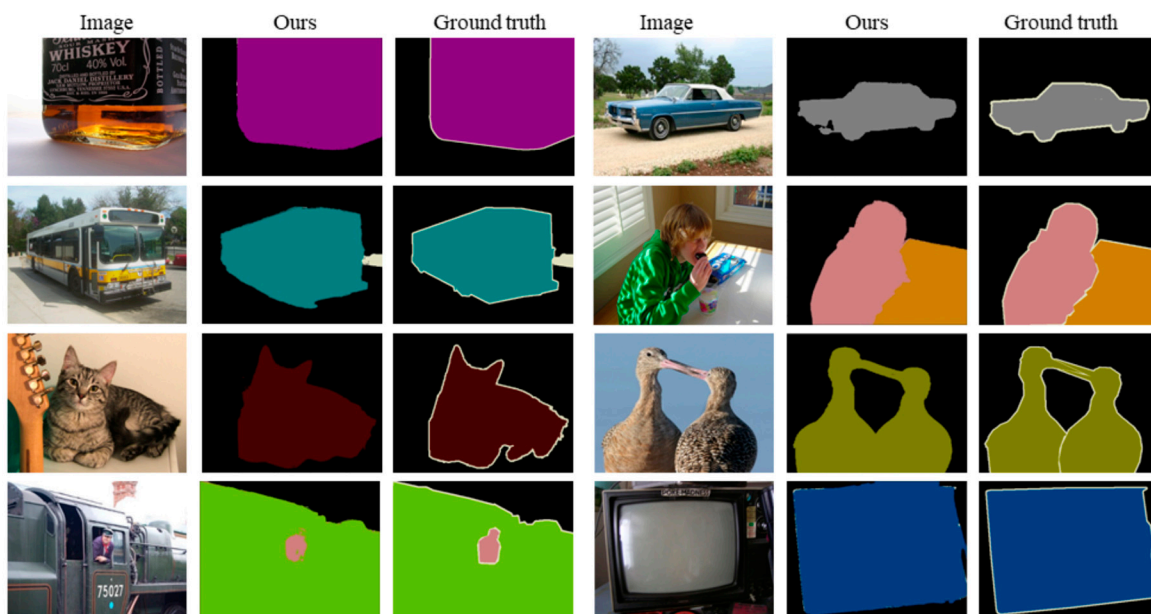

Figure 7. Some segmentation results for segmentation by our algorithm.

Table 2. Comparison with the state-of-the-art algorithms in term of IOU (%) on the 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	Stdev.
PLSA [43]	45	64	71	75	74	**86**	81	47	1	73	55	88	6	6	63	18	80	27	26	55	8	50.0	28.8
WSDC [30]	49	-	58	43	66	-	36	-	46	52	40	85	60	48	-	54	-	52	51	-	-	52.9	11.7
Textonboost [39]	**62**	**98**	**86**	58	50	83	60	53	74	63	75	63	35	19	**92**	15	**86**	54	19	**62**	7	57.8	25.5
MIM [29]	12	83	70	**81**	**93**	84	**91**	55	**97**	**87**	**92**	82	69	51	61	**59**	66	53	44	9	**58**	66.5	23.8
Ours	35	82	76	80	73	78	77	**56**	76	78	75	**89**	**72**	**60**	77	57	65	**58**	**64**	57	56	**68.6**	12.3

Table 3. Comparison with the state-of-the-art algorithms in term of IOU (%) on the PASCAL VOC 2012 dataset.

Methods	Background	Aeroplane	Bicycle	Bird	Boat	Bottle	Bus	Car	Cat	Chair	Cow	Dining table	Dog	Horse	Motorbike	Person	Potted plant	Sheep	Sofa	Train	Tv	Miou	Stdev.
CCNN [44]	69	26	18	25	20	36	47	47	48	16	38	21	44	35	46	41	30	36	22	39	37	35.3	12.6
EM-ADAPT [37]	76	37	22	42	26	38	51	45	49	17	41	**29**	47	46	55	28	30	44	**29**	34	**46**	39.6	12.9
MIL-ILP-SEG [45]	79	48	21	31	28	35	51	**55**	53	8	**56**	20	54	50	40	39	28	52	25	33	**46**	40.6	15.9
SN-B [36]	**82**	**54**	12	**54**	30	**42**	47	46	50	17	49	17	**61**	52	**62**	36	25	**58**	19	**49**	46	**43.2**	17.4
H&M [46]	78	51	18	47	32	35	**59**	51	50	16	40	27	45	48	55	32	30	44	29	34	**46**	41.9	14.2
Ours	78	35	**23**	48	**33**	39	53	**55**	52	**18**	49	24	51	**53**	44	**52**	**32**	49	24	32	43	42.2	13.8

5. Conclusions

In this paper, a novel PBCRF model is proposed for ISS with image-level labels, which provides an effective solution to the weakly supervised ISS problems. It has three advantages over existing approaches. First, based on the improved SLIC algorithm, optimal numbers of superpixels are automatically estimated for different images for improving the accuracy of image segmentation rather than using a fixed parameter. Second, by taking an image patch as the basic processing unit of ISS, this has significantly improved the performance and reduced the computational costs. Last but not the least, by combining category correlation and similarity information of each semantic category in training the PBCRF model, the inference of semantic label is transformed into the problem of minimizing a potential energy function. Extensive experimental results conducted on the MSRC and PASCAL VOC 2012 datasets segmentation benchmark have demonstrated that the proposed IPCRFWSS algorithm can produce improved or comparable results in comparison to a few state-of-the-art, even some deep learning methods. An improved or much higher mIoU along with a lower variance has also indicated the proposed approach is more robust to different semantic categories. To further improve the results in semantic image segmentation, we will focus on three topics in the future. The first is fusion of color, edge and other information for refined segmentation [47,48], and the second is saliency based extraction of objects from images [49,50]. The third direction is deep learning based image segmentation and object detection, where convolutional neural networks and other models will be explored [51,52], even in combination with the first two topics such as multiscale segmentation and extreme learning machines [53,54].

Author Contributions: Conceptualization, X.X. and Y.X.; methodology, X.X.; software, Y.X.; validation, X.X. and Y.X.; formal analysis, X.H.; investigation, Z.Z.; resources, X.X. and J.R; data curation, Y.X.; writing—original draft preparation, X.X.; writing—Review and editing X.X., Y.X. and J.R.; visualization, X.X. and Y.X.; supervision, J.X.; project administration, X.X.; funding acquisition, X.X., X.H. and J.X. All authors have read and agreed to the published version of the manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Zhang, W.; Zeng, S.; Wang, D.; Xue, X. Weakly supervised semantic segmentation for social images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 2718–2726.
2. Lin, G.; Milan, A.; Shen, C.; Reid, I. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In Proceedings of the IEEE Conference CVPR, Honolulu, HI, USA, 21–26 July 2017; pp. 1925–1934.
3. Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv Prepr* **2014**, arXiv:1412.7062.
4. Dehghani, M.; Severyn, A.; Rothe, S.; Kamps, J. Learning to learn from weak supervision by full supervision. *arXiv Prepr* **2017**, arXiv:1711.11383.
5. Wei, Y.; Xiao, H.; Shi, H.; Jie, Z.; Feng, J.; Huang, T.S. Revisiting dilated convolution: A simple approach for weakly-and semi-supervised semantic segmentation. In Proceedings of the IEEE Conference CVPR, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7268–7277.
6. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference CVPR, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
7. Lin, D.; Dai, J.; Jia, J.; He, K.; Sun, J. Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference CVPR, Las Vegas, NV, USA, 27–30 June 2016; pp. 3159–3167.
8. Pal, N.R.; Pal, S.K. A review on image segmentation techniques. *Pattern Recognit.* **1993**, *26*, 1277–1294. [CrossRef]

9. Gould, S.; Gao, T.; Koller, D. Region-based segmentation and object detection. *Adv. Neural Inf. Proc. Syst.* **2009**, *22*, 655–663.

10. Senthilkumaran, N.; Rajesh, R. Edge detection techniques for image segmentation-a survey of soft computing approaches. *Int. J. Recent Trends Eng.* **2009**, *1*, 844–846.

11. Coleman, G.B.; Andrews, H.C. Image segmentation by clustering. *Proc. IEEE* **1979**, *67*, 773–785. [CrossRef]

12. Al-Amri, S.S.; Kalyankar, N.V. others Image segmentation by using threshold techniques. *arXiv Prepr* **2010**, arXiv:1005.4020.

13. Tang, J. A color image segmentation algorithm based on region growing. In Proceedings of the 2010 2nd International Conference on Computer Engineering and Technology, Wuhan, China, 21–24 May 2010; pp. 634–637.

14. Borges, G.A.; Aldon, M.-J. A split-and-merge segmentation algorithm for line extraction in 2d range images. In Proceedings of the 15th International Conference on Pattern Recognition (ICPR), Barcelona, Spain, 3–7 September 2000; pp. 441–444.

15. Kang, W.; Yang, Q.; Liang, R. The comparative research on image segmentation algorithms. In Proceedings of the 2009 First International Workshop on Education Technology and Computer Science, Wuhan, China, 7–8 March 2009; pp. 703–707.

16. Ren, X.; Malik, J. Learning a classification model for segmentation. In Proceedings of the Ninth IEEE International Conference on Computer Vision (ICCV), Nice, France, 13–16 October 2003; pp. 10–17.

17. Shi, J.; Malik, J. Normalized cuts and image segmentation. *Dep. Pap. (CIS)* **2000**, *22*, 888–905.

18. Moore, A.P.; Prince, S.J.; Warrell, J.; Mohammed, U.; Jones, G. Superpixel lattices. In Proceedings of the IEEE Conference CVPR, Anchorage, AK, USA, 23–28 June 2008; pp. 1–8.

19. Felzenszwalb, P.F.; Huttenlocher, D.P. Efficient graph-based image segmentation. *Int. J. Comput. Vis.* **2004**, *59*, 167–181. [CrossRef]

20. Vincent, L.; Soille, P. Watersheds in digital spaces: An efficient algorithm based on immersion simulations. *IEEE Trans. Pattern Anal. Mach. Intell.* **1991**, *13*, 583–598. [CrossRef]

21. Comaniciu, D.; Meer, P. Mean shift: A robust approach toward feature space analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24*, 603–619. [CrossRef]

22. Achanta, R.; Shaji, A.; Smith, K.; Lucchi, A.; Fua, P.; Süsstrunk, S. SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *34*, 2274–2282. [CrossRef] [PubMed]

23. Hsu, C.; Ding, J. Efficient image segmentation algorithm using SLIC superpixels and boundary-focused region merging. In Proceedings of the 2013 9th International Conference on Information, Communications & Signal Processing, Tainan, Taiwan, 10–13 December 2013; pp. 1–5.

24. Ning, J.; Zhang, L.; Zhang, D.; Wu, C. Interactive image segmentation by maximal similarity based region merging. *Pattern Recognit.* **2010**, *43*, 445–456. [CrossRef]

25. Gu, X.; Deng, J.D.; Purvis, M.K. Improving superpixel-based image segmentation by incorporating color covariance matrix manifolds. In Proceedings of the 2014 IEEE International Conference on Image Processing (ICIP), Paris, France, 27–30 October 2014; pp. 4403–4406.

26. Wei, Y.; Feng, J.; Liang, X.; Cheng, M.-M.; Zhao, Y.; Yan, S. Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In Proceedings of the IEEE Conference CVPR, Honolulu, HI, USA, 21–26 July 2017; pp. 1568–1576.

27. Duygulu, P.; Barnard, K.; de Freitas, J.F.; Forsyth, D.A. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In Proceedings of the European Conference on Computer Vision (ECCV), Copenhagen, Denmark, 2002, 28–31 May; pp. 97–112.

28. Zhang, K.; Zhang, W.; Zheng, Y.; Xue, X. Sparse reconstruction for weakly supervised semantic segmentation. In Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence (IJCAI), Beijing, China, 3–9 August 2013; pp. 1889–1895.

29. Vezhnevets, A.; Ferrari, V.; Buhmann, J. Weakly supervised semantic segmentation with a Multi-Image model. In Proceedings of the IEEE Conference ICCV, Barcelona, Spain, 6–13 November 2011; pp. 643–650.

30. Liu, Y.; Liu, J.; Li, Z.; Tang, J.; Lu, H. Weakly-supervised dual clustering for image semantic segmentation. In Proceedings of the IEEE Conference CVPR, Portland, OR, USA, 23–28 June 2013; pp. 2075–2082.

31. Zhang, L.; Song, M.; Liu, Z.; Liu, X.; Bu, J.; Chen, C. Probabilistic graphlet cut: Exploiting spatial structure cue for weakly supervised image segmentation. In Proceedings of the IEEE Conference CVPR, Portland, OR, USA, 23–28 June 2013; pp. 1908–1915.

32. Wang, H.; Lu, T.; Wang, Y.; Shivakumara, P.; Tan, C.L. Weakly-supervised region annotation for understanding scene images. *Multimed. Tools Appl.* **2016**, *75*, 3027–3051. [CrossRef]

33. Noh, H.; Hong, S.; Han, B. Learning deconvolution network for semantic segmentation. In Proceedings of the IEEE Conference ICCV, Santiago, Chile, 7–13 December 2015; pp. 1520–1528.

34. Farabet, C.; Couprie, C.; Najman, L.; LeCun, Y. Learning hierarchical features for scene labeling. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *35*, 1915–1929. [CrossRef]

35. Qi, X.; Liu, Z.; Shi, J.; Zhao, H.; Jia, J. Augmented feedback in semantic segmentation under image level supervision. In Proceedings of the IEEE Conference ECCV, Amsterdam, The Netherlands, 11–14 October 2016; pp. 90–105.

36. Wei, Y.; Liang, X.; Chen, Y.; Jie, Z.; Xiao, Y.; Zhao, Y.; Yan, S. Learning to segment with image-level annotations. *Pattern Recognit.* **2016**, *59*, 234–244. [CrossRef]

37. Papandreou, G.; Chen, L.-C.; Murphy, K.; Yuille, A. Weakly-and semi-supervised learning of a DCNN for semantic image segmentation. *arXiv Prepr* **2015**, arXiv:1502.02734.

38. Wei, Y.; Liang, X.; Chen, Y.; Shen, X.; Cheng, M.-M.; Feng, J.; Zhao, Y.; Yan, S. Stc: A simple to complex framework for weakly-supervised semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *39*, 2314–2320. [CrossRef]

39. Shotton, J.; Winn, J.; Rother, C.; Criminisi, A. Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. In Proceedings of the IEEE Conference ECCV, Graz, Austria, 7–13 May 2006; pp. 1–15.

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

41. Everingham, M.; Eslami, S.A.; Van Gool, L.; Williams, C.K.; Winn, J.; Zisserman, A. The pascal visual object classes challenge: A retrospective. *Int. J. Comput. Vis.* **2015**, *111*, 98–136. [CrossRef]

42. Liu, M.-Y.; Tuzel, O.; Ramalingam, S.; Chellappa, R. Entropy rate superpixel segmentation. In Proceedings of the IEEE Conference CVPR, Providence, RI, USA, 20–25 June 2011; pp. 2097–2104.

43. Verbeek, J.; Triggs, B. Region classification with markov field aspect models. In Proceedings of the IEEE Conference CVPR, Minneapolis, MN, USA, 17–22 June 2007; pp. 1–8.

44. Pathak, D.; Krahenbuhl, P.; Darrell, T. Constrained convolutional neural networks for weakly supervised segmentation. In Proceedings of the IEEE Conference ICCV, Santiago, Chile, 7–13 December 2015; pp. 1796–1804.

45. Pinheiro, P.O.; Collobert, R. From Image-Level to Pixel-Level Labeling with Convolutional Networks. In Proceedings of the IEEE Conference CVPR, Boston, MA, USA, 7–12 June 2015; pp. 1713–1721.

46. Redondo-Cabrera, C.; Baptista-Ríos, M.; López-Sastre, R.J. Learning to exploit the prior network knowledge for weakly supervised semantic segmentation. *IEEE Trans. Image Process.* **2019**, *28*, 3649–3661. [CrossRef] [PubMed]

47. Yan, Y.; Ren, J.; Li, Y.; Windmill, J.F.; Ijomah, W.; Chao, K.-M. Adaptive fusion of color and spatial features for noise-robust retrieval of colored logo and trademark images. *Multidimens. Syst. Signal Process.* **2016**, *27*, 945–968. [CrossRef]

48. Xie, X.; Xie, G.; Xu, X.; Cui, L.; Ren, J. Automatic image segmentation with superpixels and image-level labels. *IEEE Access* **2019**, *7*, 10999–11009. [CrossRef]

49. Wang, Z.; Ren, J.; Zhang, D.; Sun, M.; Jiang, J. A deep-learning based feature hybrid framework for spatiotemporal saliency detection inside videos. *Neurocomputing* **2018**, *287*, 68–83. [CrossRef]

50. Yan, Y.; Ren, J.; Sun, G.; Zhao, H.; Han, J.; Li, X.; Marshall, S.; Zhan, J. Unsupervised image saliency detection with Gestalt-laws guided optimization and visual attention based refinement. *Pattern Recognit.* **2018**, *79*, 65–78. [CrossRef]

51. Han, J.; Zhang, D.; Hu, X.; Guo, L.; Ren, J.; Wu, F. Background prior-based salient object detection via deep reconstruction residual. *IEEE Trans. Circuits Syst. Video Technol.* **2014**, *25*, 1309–1321.

52. Han, J.; Zhang, D.; Cheng, G.; Guo, L.; Ren, J. Object detection in optical remote sensing images based on weakly supervised learning and high-level feature learning. *IEEE Trans. Geosci. Remote Sens.* **2014**, *53*, 3325–3337. [CrossRef]

53. Xu, X.; Li, G.; Xie, G.; Ren, J.; Xie, X. Weakly supervised deep semantic segmentation using CNN and ELM with semantic candidate regions. *Complexity* **2019**, *2019*, 1–12. [CrossRef]

54. Huang, H.; Sun, G.; Zhang, X.; Hao, Y.; Zhang, A.; Ren, J.; Ma, H. Combined multiscale segmentation convolutional neural network for rapid damage mapping from postearthquake very high-resolution images. *J. Appl. Remote Sens.* **2019**, *13*, 1–14. [CrossRef]