Journal of Information Hiding and Multimedia Signal Processing ©2019 ISSN 2073-4212
Ubiquitous International Volume **10**, Number **1**, January **2019**

Digital Image Semantic Segmentation Algorithms: A Survey

Wei Song[1,*], Nan Zheng[1], Rui Zheng[1], Xiao-Bing Zhao[1], Antai Wang[2]

[1]School of Information and Engineering
Minzu University of China
Beijing, 100081, P. R. China
[2]Department of Mathematical Sciences
New Jersey Institute of Technology
Newark, NJ, USA

songwei@muc.edu.cn

Received June 2018; revised October 2018

Abstract. *In the field of computer vision, image semantic segmentation is an important research branch and it is also a challenging task. Applications such as autonomous driving, Unmanned Aerial Vehicle System (UAVS), and even virtual or augmented reality systems require accurate and efficient segmentation mechanisms. With the rise of deep learning methods, image semantic segmentation is more and more concerned by relevant researchers. In order to understand the research status, existing problems and development prospects of image semantic segmentation, this paper introduces the mainstream image semantic segmentation methods on the basis of extensive survey. First of all, we introduce the background concept of image semantic segmentation, generalize the commonly used image semantic segmentation methods, and compare the segmentation results of each method. After that, the commonly used image semantic segmentation datasets are summarized. At the same time, several commonly evaluation standards are introduced. Finally, the future development trend of image semantic segmentation is prospected, with a view to providing some ideas for researchers who wish to engage in this field.*

Keywords: Image semantic segmentation, Neural network, Deep learning.

1. **Introduction.** Image segmentation [1] refers to dividing an image into several non-overlapping regions based on basic image features such as grayscale, color, texture, and shape, and the segmentation method makes these features appear similar in the same region and different between different regions. The segmentation results can be used to extract the target of interest. The early image segmentation is usually classified into two categories: the one is foreground, the other is background. With the complication of images, people are eager to classify and recognize specific objects in the image, and thus semantic segmentation of images gradually arises.

According to the degree of segmentation, from coarse to fine, it can be divided into image classification, object detection, image segmentation and instance segmentation. Image classification is for the entire image, marking the objects contained in the image. Object detection is to mark the position of the corresponding object in the image, and semantic segmentation is to accurately segment and mark the object in the image, higher level of segmentation is to mark each object, and the objects of the same category are also marked separately, that is, instance segmentation.

This paper firstly introduces the traditional methods of image semantic segmentation and neural network segmentation methods in recent years, and compares the experimental results of these methods discussed. Secondly, we list the commonly used standard semantic segmentation datasets and the basic information of these datasets; and then, the detail of the evaluation criteria is given. Finally, the future development direction of image semantic segmentation is predicted.

2. Overview on image semantic segmentation.

In general, we hope that the machine can automatically segment and recognize objects existing in the image. The traditional image semantic segmentation technology is mainly based on spectral clustering theory. According to the relationship weights between different pixels, the image is divided into two categories according to the given threshold value. This simple and crude segmentation method results in inaccurate segmentation results. Although it has been improved based on the map-cutting theory, the segmentation process requires human intervention and is not suitable for rapid batch processing. With the continuous improvement of deep neural network algorithm, the semantic segmentation of images has been further developed, the features have been extracted quickly and accurately and the segmentation results are more accurate.

2.1. Traditional image semantic segmentation methods.

Shi et al. [2] proposed the normalized cut (N-Cut) algorithm in 2000. Compared with the traditional minimization cut criterion, the normalized cut criterion not only satisfies the minimum similarity between classes, but also satisfies the maximum similarity within class. Which is defined as follows:

$$Ncut(A,B) = \frac{cut(A,B)}{assoc(A,V)} + \frac{cut(A,B)}{assoc(B,V)} \tag{1}$$

Where, the cut is defined below:

$$cut(A,B) = \sum_{u\in A, v\in B} w(u,v) \tag{2}$$

A and B are two disjoint sets in graph $G = (V, E)$, where $A \cup B = V$, $A \cap B = \emptyset$, $w(u, v)$ is the similarity function between u and v nodes. $assoc(A, V) = \sum_{u\in A,t\in V} w(u, t)$ represents the sum of similarity functions between u and t nodes, u denotes all nodes in A, and t denotes all nodes in the graph. However, in practical applications, the N-Cut algorithm can only perform the cutting of the image once per execution, and thus it is inevitable when the image is multi-objects, the algorithm needs to be executed successively many times, which will result in inaccurate segmentation results. Zhao et al. [3] proposed a fast image segmentation method that used a simple linear iterative clustering super-pixel algorithm to obtain super-pixel regions, and then affinity propagation clustering was used to extract the representative pixels of each super-pixel region, eventually uses N-cut to get segmentation results. Boykov [4] proposed the Graph Cuts algorithm in 2001. This algorithm took into account the energy with various smoothing constraints. It not only used the pixel gray information of the image, but also considered the regional boundary information, and combined the global optimum, which guaranteed the segmentation effect.

Wen [5] proposed an image segmentation algorithm based on weakly supervised learning and secondary clustering. This algorithm combined spectral clustering and discriminative clustering. First, it uses spectral clustering to learn the category indicator function, which would guide discriminative clustering to learn potential data features. The algorithm was robust and the performance was good on untagged images. Liu et al. [6] proposed a weakly supervised dual clustering method for image level label. The image semantic segmentation method uses spectral clustering to cluster super-pixels obtained from a set of over-segmented images. At the same time, the linear transformation between features and annotations is used as a discriminative clustering learning. In order to select the discriminative features between different categories, the two clustering results should be as consistent as possible. Finally, the iterative CCCP procedure is used to optimize non-convex and non-smooth objective functions. Guo [7] proposed an improved algorithm for the noise sensitivity problem of the traditional fuzzy C-means clustering algorithm. This algorithm uses the surrounding environmental information of the reference image to join the grouping calculations to reduce the influence of noise, as well as speeding up the running time of the FCM clustering algorithm. Its minimum objective function is defined as in

$$J = \sum_{i=1}^{c} \sum_{k=1}^{t} \mu_{i,k}^{m} \|x_k - v_i\|^2 + \alpha \sum_{i=1}^{c} \sum_{k=1}^{t} \mu_{i,k}^{m} \|\bar{x}_k - v_i\|^2 \tag{3}$$

Where the $\bar{x}_k$ can be taken as the average value of all nearby points around x_k and its own. The $u_{i,k}$ and v_i are defined as

$$\mu_{i,k} = \frac{(\|x_k - v_i\|^2 + \alpha\|\bar{x}_k - v_i\|^2)^{\frac{-1}{m-1}}}{\sum_{j=1}^{c}(\|x_k - v_j\|^2 + \alpha\|(\|x_k - v_i\|^2 + \alpha\|\bar{x}_k - v_i\|^2)^{\frac{-1}{m-1-v_j}}\|^2)^{\frac{-1}{m-1}}} \tag{4}$$

$$v_i = \frac{\sum_{k=1}^{t} \mu_{i,k}^{m}(x_k + \alpha\bar{x}_k)}{(1+\alpha)\sum_{k=1}^{t} \mu_{i,k}^{m}} \tag{5}$$

The traditional image semantic segmentation algorithms based on clustering method are shown in Table 1.

Table 1. Comparison of algorithms based on clustering methods (%)

Author	Algorithm features	Datasets	Segmentation results
WEN[5]	Weak supervision,spectral clustering,discriminative clustering	MSRC-21	70(mA)
LIU[6]	Weak supervision, double-end clustering	MSRC-21	52.9($mIoU$)
		LABLEME	26(mA)
GUO[7]	FCM algorithm, grouping algorithm	Self-built dataset	2.2($mError$)

Zhang [8] proposed an image semantic segmentation method based on the probability map model, using a traditional high-order conditional random field model to establish a model which was based on different degrees of quantification images. At the same time, in order to strengthen the robustness of the algorithm, feature variance was used to determine the pixels of inconsistent labels. Zhang et al. [9] proposed a discriminant model based on conditional random field to learn the posterior conditional probability distribution of different class labels, and then maximized the posterior probability to obtain the best label. The algorithm used a shape filter to describe the underlying texture features of the image and the context information based on the surrounding pixel's texture features. Zuo et al. [10] proposed a RGB-D image semantic segmentation method based on the interactive conditional random field, which could be effectively applied to complex and varied real world scenes. In this paper, the morphological reconstruction methods were used to preprocess the image aiming at reducing the image noise and data loss, then the conditional random field method was used to roughly segment the image, and finally improving the segmentation result through a human-computer interaction platform. Its energy term is defined as

$$E(c) = \sum_{i} E_1(c_i : x_i) + \lambda \sum_{i,j} E_2(c_i : c_j) \tag{6}$$

Where, $E_1(c_i : x_i)$ measures the probability that the pixel i is labelled c_i under feature x_i, $E_2(c_i : c_j)$ measures the consistency of two connected pixels' label.

In [11] a new high-order conditional random field was proposed. The model combined the target detection results based on global shape features and the point-to-condition model. Target detectors and pre-background segmentation algorithms were used to obtain target regions in the image, and new high-level energy items were defined on the target regions. The new high-order conditional random field model was a weighted mixed model of high-order energy items and point-pair conditional random field models, its optimal solution was the final semantic segmentation result of the image. The new high-order energy term is defined as:

$$E_{d_k}(x_{d_k}) = -|x_{d_k}|max(0,(1-R)max(0,(C_{d_k}-C_i))) \tag{7}$$

$$R = \frac{N_{d_k}}{R_t|x_{d_k}|} \tag{8}$$

Where x_{d_k} is a set of random flag variables corresponding to all pixels that make up a single object area, C_t is the threshold. By adjusting this value, the final recognition accuracy rate can be controlled. Wang et al. [12] proposed an improved image segmentation algorithm based on a robust high-order conditional random field model, according to the given tag set, the maximum stream-minimum cut algorithm was applied to obtain the local optimal solution, then the local optimal solution was used to modify the node's tag, and the extended algorithm was run on the unmarked nodes. At the same time, the flow and edge of the graph were dynamically updated during each iteration, which would make the time of each iteration decrease rapidly. The experimental results showed that the convergence speed was faster on the same segmentation effect. The image semantic segmentation algorithms based on conditional random field are shown in Table 2.

Table 2. Comparison of algorithms based on conditional random field (%)

Author	Algorithm features	Datasets	Segmentation results
ZHANG[8]	CRF, dense features, high-order potential energy	MSRC-21	75.8(mA)
ZHANG[9]	CRF, Joint-boosting Algorithm	MSRC-21	71.6(mA)
ZUO[10]	CRF, Interactive	Self-built dataset	95.3(mA)
MAO[11]	CRF, high order energy items	MSRC-21	72.2(PA)
WANG[12]	CRF, Maximum Flow - Minimum Cut	MSRC-21	0.7s(time)

Chen et al. [13] proposed a new image semantic segmentation model in combination with the underlying segmentation results. First, the corresponding underlying segmentation image block was obtained by the histogram threshold and the K-means. Then the high-level semantic information of the image was acquired by the word bag model. Finally, the high-level semantic information was used in conjunction with the support vector machine re-labels the image block to obtain the final image semantic segmentation result. In [14] an image semantic segmentation algorithm based on texture primitive blocks was proposed. Firstly, texture primitive features were extracted, k-means and k-d trees were used to get the image's texture primitive block segmentation maps, and then semantic mapping of texture primitive blocks was implemented by using the image semantic learning and prediction methods based on support vector machine.

The two papers have similar ideas. Firstly, the image is subdivided and then the high-level semantic information of the image is obtained. Then the support vector machine is

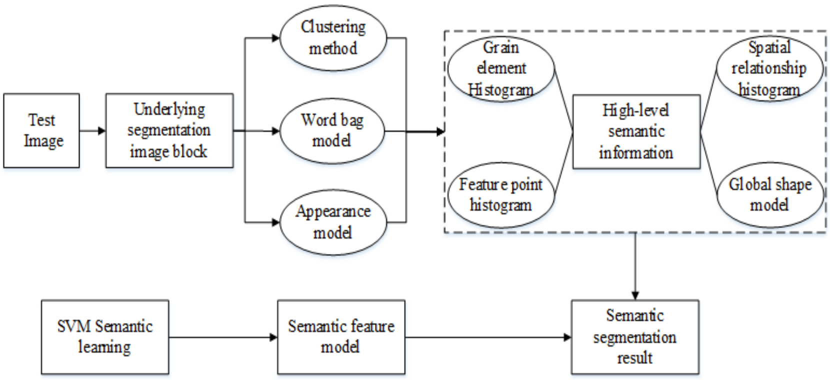

Figure 1. Flowchart of the proposed method

combined to optimize the image segmentation result and the final semantic segmentation result is obtained, the framework is shown in Figure 1 .

Zhang [15] proposed an image semantic segmentation algorithm based on random forest, randomly sampling a fixed-size window as a feature from a training image, and quantizing these features as numeric vectors by comparing the pixel values of two pixels at random in the window. These vectors were used to train the random forest classifier. When we test the classifier, each pixel was taken as the center, based on which, a window was extracted, and then a vector set was extracted in the window. Finally these vectors were respectively performed voting using the leaf nodes of the random forest. The most likely attribution category of this pixel was selected based on the voting result. Because the algorithm used low-level pixel information of the image, the algorithm's training and test speed had been greatly improved. Cao et al. [16] proposed an image semantic segmentation method based on image hierarchical tree. The hierarchical tree model used the structured forest method to generate the contour model. Considering the possible over-segmentation problem, the multi-scale contour map algorithm was used to obtain the multi-scale contour map. Then the multi-scale contour map was trained using the support vector machine to generate the image hierarchy tree. The final output image semantic segmentation results were obtained by refining hierarchical tree through random forest.

2.2. **Neural network semantic segmentation methods.** Most of the traditional semantic segmentation methods are based on the low-order visual information of the image itself. Therefore, in the more difficult segmentation tasks, such as the need for artificial auxiliary information, such segmentation results are often not ideal. Semantic segmentation of images has always been a part of computer vision technology since 2007, but as other fields of computer vision technology, image semantic segmentation has achieved a major breakthrough when Long et al. [17] used the fully convolutional neural network in 2014.

In 2014, Long et al. [17] proposed the concept of fully convolutional networks (FCN), aiming to produce output of corresponding size with arbitrary size input and effective reasoning. The traditional network structure has been improved and the learning representation has been applied to the segmentation task, which has increased the mIoU of the PASCAL VOC2012 dataset by nearly 0.2 . Figure 2 shows the FCN test results on common semantic segmentation datasets, such as PASCAL VOC2012, SIFT Flow and so on.

In the traditional methods, we should semantically segment the image to generate different regions on the image first, then extract features from the regions, and then combine the regions to get the final result of semantic segmentation. Chen et al. [18] learned the

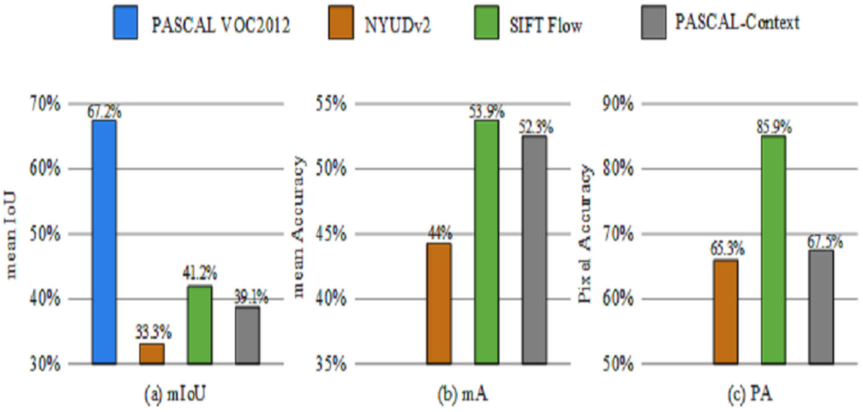

Figure 2. FCN Results on common semantic segmentation datasets

process was complex and the effect could be improved, constructed and implemented a deep neural network structure that combines a convolutional neural network and a deconvolution layer neural network, and directly predicted the semantic category at the pixel level. The optimized algorithm had achieved good results. Noh et al. [19] proposed a new image semantic segmentation method by learning the deep deconvolution network. The network consisted of deconvolution and decapitalization layers, where the deconvolution layer determines pixel-level class labels and the pooling layer predicted the segmentation mask. Then the algorithm used the trained network for each resolution of the input image, and finally combined all the results to form the final semantic segmentation map. Considering the network integrated the deep deconvolution network and the proposed forecasting, it alleviated to some extent the existing limitations of FCN.

Zhao [20] proposed PSPNet, which incorporated global feature information, and had modified the underlying ResNet architecture in conjunction with an extended convolution. The algorithm introduced an auxiliary loss in the middle layer of ResNet, which had made the entire learning process optimized in the public datasets such as CityScapes, ADE20K, etc. Considering the problem of repeated subsampling in deep convolutional neural networks, a multipath optimization network, RefineNet was proposed by Lin et al. [21]. All the information available in the downsample process was utilized, and high resolution prediction was achieved in combination with long-range residual connections. In addition, chained residual pools were introduced to capture rich background content in a more efficient manner. The network iteratively increased the resolution of the feature over several resolution ranges using a special RefineNet block and ultimately produced a high resolution segmentation map.

Hong et al. [22] used heterogeneous annotations to propose a novel deep neural network to solve the semi-supervised semantic segmentation problem. In this architecture, the classification network was used to identify the tags associated with the images. The decoupled architecture ensured that the classification and segmentation networks could be learned separately from training data labeled at the image level and at the pixel level, helping to effectively reduce the use of specific classes of activation maps obtained from the bridge layer. The search space used for segmentation presented a great advantage compared to other semi-supervised methods. The image semantic segmentation algorithms based on the PASCAL VOC dataset are shown in Table 3.

Vijay et al. [23] used SegNet, an architecture of deep convolutional coding and decoding. The architecture of the coding network was topologically identical to the 13 convolutional layers in the VGG-16 network. The role of the decoder network was to map

Table 3. Comparison of algorithms based on PASCAL VOC2012 dataset (%)

Author	Algorithm features	Datasets	Segmentation results
CHEN[18]	Convolution and Deconvolution Neural Networks	PASCAL VOC2012	63.6(mIoU)
NOH[19]	Deconvolution Networks	PASCAL VOC2012	72.5(mIoU)
ZHAO[20]	PSPNet	PASCAL VOC2012	82.6(mIoU)
LIN[21]	RefineNet	PASCAL VOC2012	83.4(mIoU)
LONG[22]	Decoupled Deep Neural Networks	PASCAL VOC2012	66.6(mIoU)

the low resolution encoder feature map to a full input resolution feature map for pixel level classification. The novelty of SegNet was that the decoder performs non-linear upsample using the convergence index calculated in the maximum pooling step of the corresponding encoder, thus eliminating the need for learning upsample and achieving a balance between memory and precision. Simon et al. [24] extended DenseNets to handle semantic segmentation of images. It fed each layer to each layer in a feed-forward manner. For each layer, the feature map of all previous layers was used as input, and its own feature map was used as input for all subsequent layers. The network reduced the gradient disappearance problem, features were reused, and the number of parameters used was greatly reduced. A good segmentation result was obtained on CamVid and other datasets.

Adam [25] considered image semantics to be segmented in real time. However, traditional neural networks required a large number of floating-point operations and run for a long time. Naturally, the ENet structure was proposed to specifically serve tasks that required low-latency operations. Islam et al. [26] proposed a G-FRNet architecture for the coding-decoding architecture to obtain higher frequency details in the elaboration stage. This was an end-to-end dense marking task deep learning framework. The framework addressed the limitations of existing methods. The network firstly performed a rough prediction, and then gradually refined the details by effectively integrating the local and global context information during the refinement phase. Finally, a gating unit was introduced to control the information passed to filter out ambiguity. Nasim et al. [27] proposed a semi-supervised framework based on generative confrontation network (GAN). The basic idea was to add a large amount of virtual visual data, forcing the real samples to approach each other in the feature space, thereby achieving bottom-up clustering. The process further improved multi-class pixel classification. The framework consisted of a network of generators, provided additional training examples for multiple classes of classifiers, and added weakly annotated data to extend the above framework. The algorithm had yielded good results. Image semantic segmentation algorithms based on CamVid dataset are shown in Table 4.

The quantitative comparisons on the CamVid dataset which aims to segment 11 road classes are showed in Figure 3.

He et al. [28] extended Faster R-CNN, proposed Mask R-CNN, added a branch for predicting the object mask, and combined it with the existing method for boundary box recognition. This algorithm was faster and easier to generalize to other tasks. Li et al. [29] proposed the first end-to-end fully convolutional neural network to perform real-level image semantic segmentation, by introducing position-sensitive internal and external score maps, basic convolutional representation could be shared between two subtasks. Experiments have shown that the network had good performance in accuracy and efficiency. Dai et al. [30] proposed a multi-tasking network cascading instance-aware

Table 4. Comparison of algorithms based on CamVid dataset (%)

Author	Algorithm features	Datasets	Segmentation results
VIJAY[23]	SegNet	CamVid	60.1(mIoU)
SIMON[24]	Densely Connected Convolutional Networks	CamVid	66.9(mIoU)
ADAM[25]	ENet	CamVid	51.3(mIoU)
ISLAM[26]	Gated Feedback Refinement Networks	CamVid	68.0(mIoU)
NAISM[27]	Generative Adversarial Networks	CamVid	58.2(mIoU)

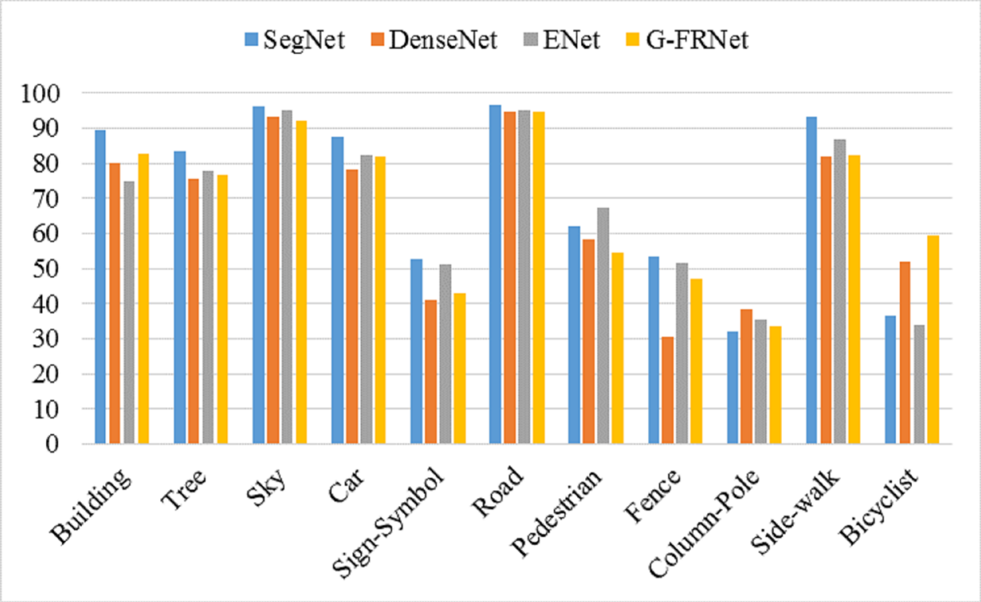

Figure 3. Quantitative comparisons on CamVid dataset

semantic segmentation network, which consisted of three parts: instance discrimination, mask estimation, and object classification. These networks formed a cascade structure and could share convolution features. A great improvement had been made in algorithm time. He et al. [31] proposed a residual learning framework to alleviate the training problem of the deeper network structure. The input layer learned the residual function instead of the unreferenced function. The experimental results proved the network was easier to optimize. The image semantic segmentation algorithms based on MS COCO dataset are shown in Table 5.

Table 5. Comparison of algorithms based on MS COCO dataset (%)

Author	Algorithm features	Datasets	Segmentation results
HE[28]	Mask R-CNN	MS COCO	37.1(PA)
LI[29]	FCIS	MS COCO	59.9(PA)
DAI[30]	Multitask Network Cascades	MS COCO	51.5(PA)
HE[31]	Residual Networks	MS COCO	48.4(PA)

Olaf [32] constructed a U-shaped network. Successful training of deep networks requires a large number of annotated training samples. U-shaped network relied on powerful data enhancement techniques to achieve more efficient use of available annotated samples. The image was trained end to end. The result of the segmentation of the neuron structure was better than the optimal sliding window convolutional network. U-Net was a typical semantic segmentation model based on the encoder-decoder structure. The encoder part was designed to determine what the object is, and the decoder part was to determine the pixel position, so that the image was segmented, usually end-to-end segmentation. U-Net structure was small and elegant and applied to small number datasets. U-Net structure is shown in Figure 4.

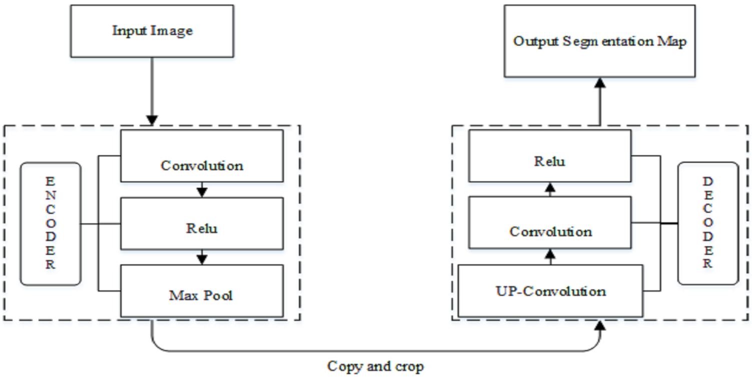

Figure 4. U-Net framework model

Liu [33] addressed the problem that most image semantic segmentation methods need to manually design image features, and used the convolutional neural network to automatically learn the advantages of image features, and comprehensively considered the CNN's network input and object context relationships for image semantic segmentation results. With the influence of super-pixels as the basic processing unit, a multi-scale CNN model for image semantic segmentation was designed in combination with multi-scale techniques.

3. **Image semantic segmentation datasets.** In order to more accurately and consistently compare the semantic segmentation results of images, standard image datasets are needed for evaluation. Common datasets include BSDS500, SUN RGB-D, PASCAL VOC2012, CITYSCAPES, MSRC-21, and MS COCO. Their comparison is shown in Table 6, including number of images, number of classes and size of images.

Table 6. Comparison of main image segmentation datasets

Dataset	Number of images	Number of classes	Size of images/pixel
BSDS500	500	–	481 × 321
SUN RGB-D	13215	19	561 × 427
PASCAL VOC2012	9993	21	500 × 375
CITYSCAPES	25000	30	adaptive
MSRC-21	591	21	320 × 213
MS COCO	328124	91	adaptive

The six datasets are separately introduced as following, describing in detail the specifics of the datasets and showing some examples of the image.

BSDS500 [34] whose full name is Berkeley Segmentation Dataset and Benchmarks 500, which is provided by the Computer Vision Group of Berkeley University. The dataset contains 500 natural images, of which 200 training image sets, 100 verification image sets and 200 test image sets. The ground truth of the image is manually identified, and the image number in this dataset is used as a unit. At the same time, the file is saved in the mat format. One file contains tag information of multiple taggers. The segmentation effect of the algorithm is evaluated by Recall and Precision. Some examples of this dataset are shown in Figure 5.

Figure 5. Examples of BSDS500 dataset

The SUN RGB-D [35] dataset contains image color information and distance information. The images are all indoor scenes and contain 300 common home objects. These objects are arranged in 51 categories according to the WordNet hypernym-hyponym relationship, 19 of which can be used for image segmentation tasks. In addition to the isolated views of 300 objects, the RGB-D object dataset also includes 22 annotated video sequences. At present, the dataset has 10355 training images and corresponding truth information, 2860 newly collected test images and corresponding metadata. Some examples of this dataset are shown in Figure 6.

Figure 6. Examples of SUN RGB-D dataset

The PASCAL VOC2012 [36] dataset provides a set of standardized datasets for object segmentation, object detection and object classification. The PASCAL VOC2012 dataset is spawned from the PASCAL VOC Challenge to test the pixel-level segmentation on an image and define each pixel which category it belongs to, including background class. At present, the dataset contains 20 object categories and a background category, involving categories such as person, bird, cat, boat, bus, and so on. There are 9993 images used to the task of segmenting, of which 2913 images are training sets involving in 6929 objects, collecting from 2007 to 2011, while the test sets only contain all the pictures from 2008 to 2011. The image pixel size is about 500 × 375 or so, the deviation does not exceed 100 pixels. Some examples of images are shown in Figure 7.

When the PASCAL VOC2012 dataset is used for segmentation tasks, the image-specific class segmentation labels and object segmentation labels are respectively given. Some of the segmentation labels are shown in Figure 8. The first column is the original image, the second column is the class segmentation mask image corresponding to the original

Figure 7. Examples of PASCAL VOC2012 dataset

image, and the third column is the object segmentation mask image corresponding to the original image.

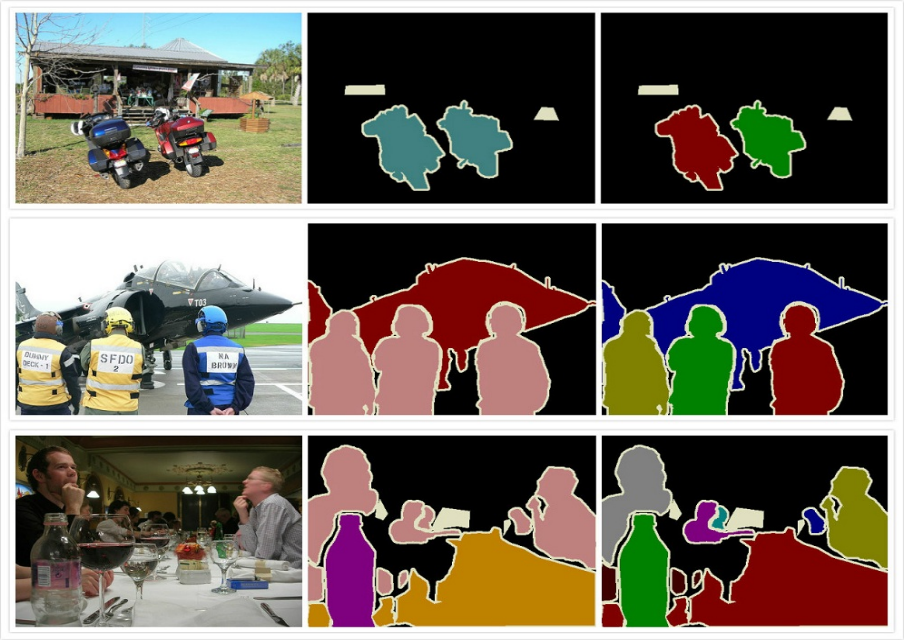

Figure 8. Examples of PASCAL VOC2012 groundtruth

The CITYSCAPES [37] dataset collects street scenes of different scenes, different backgrounds, and seasons in 50 cities. There are 30 categories involving road, person, pole, and so on. The dataset provides two versions of annotated images, with 5,000 fine-labeled images and 20,000 rough-labeled images. It is a good dataset for pixel-level semantic label segmentation and instance-level semantic label segmentation. The dataset uses IU as the evaluation index to measure the segmentation effect. Some examples of images are shown in Figure 9.

Figure 10 shows examples of CITYSCAPES dataset groundtruth. The first column is the original image, the second column is the coarse mask image corresponding to the original image, and the third column shows the fine mask image corresponding to the original image.

The MS COCO [38] dataset is acquired by the Microsoft team and mainly captured from complex everyday scenes. This dataset involves 91 categories of targets, including 165,482 training images, 81,208 validation images, and 81,434 test images. Some image examples are shown in Figure 11.

Figure 9. Examples of CITYSCAPES dataset

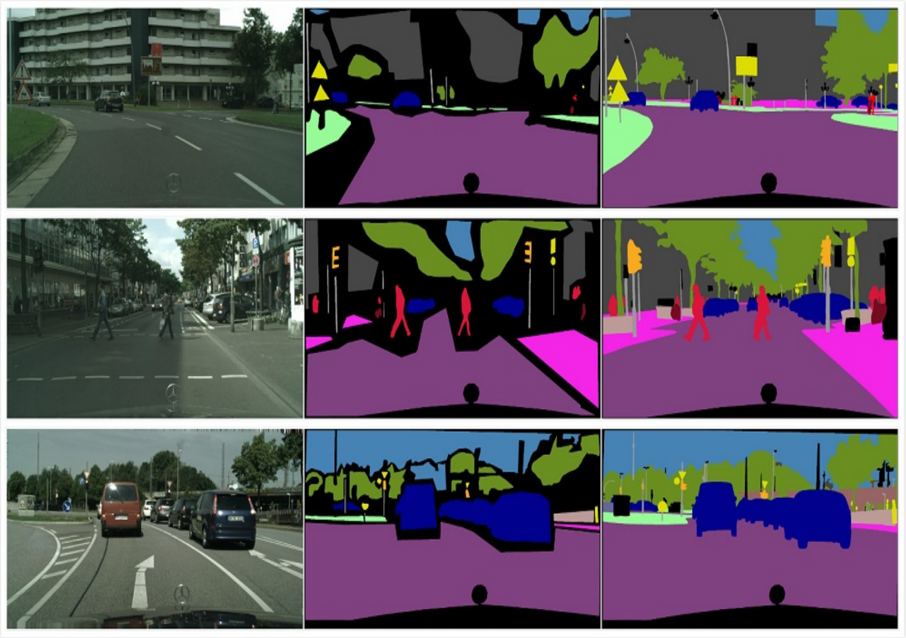

Figure 10. Examples of CITYSCAPES groundtruth

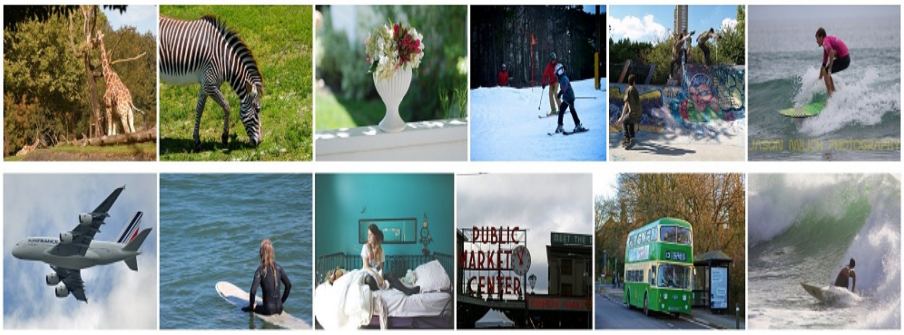

Figure 11. Examples of MS COCO dataset

The MSRC-21 [39] dataset was published by the Microsoft Cambridge Research Center and is one of the most complex and hand-labeled databases. This dataset consists of 591 images and contains 23 types of objects, of which the images of horses and mountains occupy a small part. Different colors in the labeling diagram represent different types of objects, and the black represents empty categories. A partial image examples are shown in Figure 12.

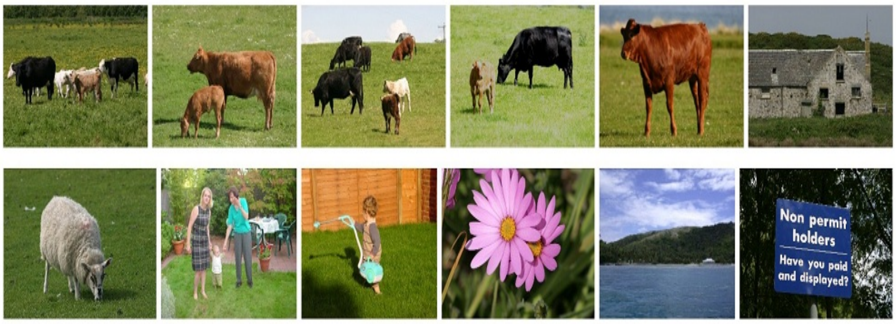

Figure 12. Examples of MSRC-21 dataset

The MSRC-21 dataset are not only some coarse mask images but also some high quality annotations. Some annotation examples of this dataset are shown in Figure 13 below. The first row shows the results of the image and its corresponding coarse annotation. The second row shows the images and their corresponding high-quality annotation results.

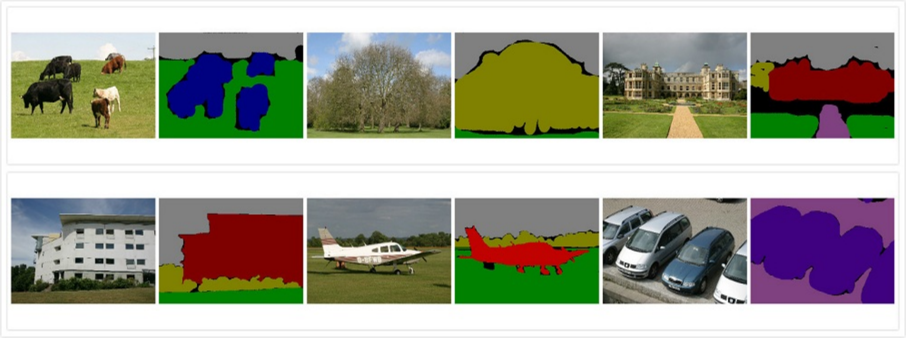

Figure 13. Examples of MSRC-21 groundtruth

With the continuous development of image semantic segmentation technology, more and more datasets for image segmentation are also being disclosed. Especially since 2015, the number of datasets has increased significantly. Figure 14 summarizes the datasets that have been publicized since 2000.

4. **Evaluation criteria.** In order to objectively and scientifically evaluate the performance of semantic segmentation algorithms, it is necessary to use quantitative methods to calculate the performance indicators of the segmented images. The measurement indices generally include PA, MA, and mean IU, as shown in Equations 9, 10 and 11 respectively. Let N_{ij} denote the number of pixels that the category i is predicted as category j; n_c represents the total number of different categories.

PA: Pixel Accuracy, refers to the ratio of the correct pixels to the total number of pixels in the segmentation result. It is usually defined as in

$$PA = \frac{\sum_i N_{ii}}{\sum_i \sum_j N_{ij}} \tag{9}$$

mA: Mean Accuracy is the average accuracy rate, which refers to the average of the accuracy of all categories of pixels in the dataset used. It is usually defined as in

$$mA = \frac{1}{n_c} \times \sum_i \frac{N_{ii}}{\sum_j N_{ij}} \tag{10}$$

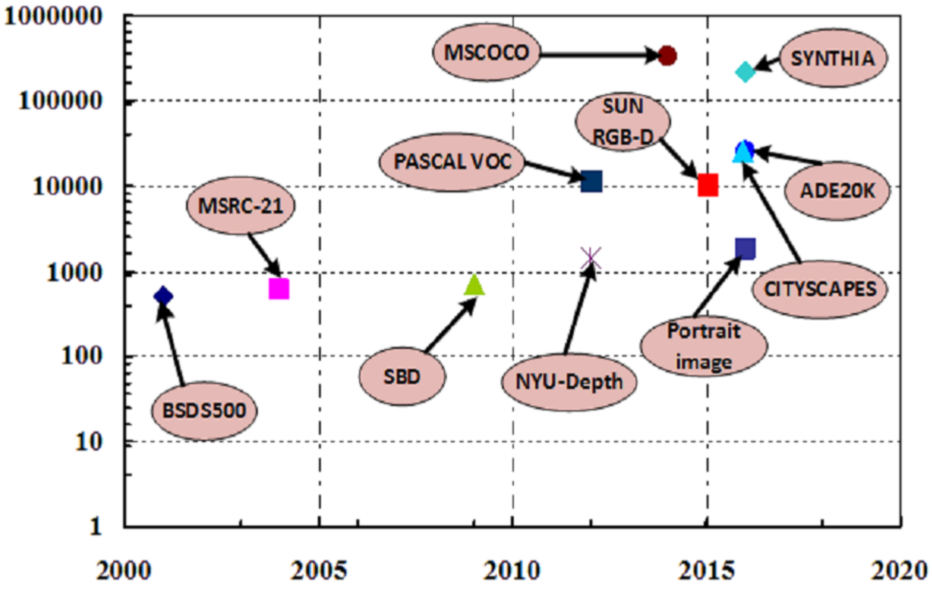

Figure 14. Increasing dataset size over time since 2000. The abscissa represents the year, and the ordinate represents the number of samples contained in the dataset.

mIoU: mIoU refers to the average of the ratio of the intersection of the segmentation result of all categories in the dataset with the true value and the union of the segmentation result and the true value, usually defined as in

$$mIoU = \frac{1}{n_c} \times \sum_i \frac{N_{ii}}{\sum_j N_{ij} + \sum_j N_{ji} - N_{ii}} \tag{11}$$

5. **Conclusion.** Image semantic segmentation isn't an isolated field. Image preprocessing and enhancement are all helpful to image segmentation. Image semantic segmentation can promote the task of image understanding. With the advent of the artificial intelligence wave, deep learning technology has also made great achievements in the semantic segmentation of images. On the basis of the fully convolutional neural network, many scholars continue to try to optimize the network, learn the hierarchical features, and make great success. However, there are also some deficiencies, such as the lack of perception of different features, which limits the application of segmentation technology in specific problems or scenes; spatial invariance leads to the inability to consider global context information and perform real-time processing on high-resolution images as well as real-time processing speed can't be reached and so on. These require us to constantly improve existing technologies and promote the development of image segmentation.

The current method of machine learning still belongs to supervised learning. The future research direction can achieve semi-supervised or weakly supervised learning, and it is closer to the cognitive style of human beings. It will also have a profound impact on semantic segmentation at the instance level. Taking into account the time-consuming problem of training model, we can consider using GPU acceleration, cloud computing and other methods to speed up the realization of image semantic segmentation, which will help achieve the goal of faster and more accurate segmentation.

Acknowledgment. The financial support and general support for this survey provided by the National Natural Science Foundation of China (No. 61503424, 61331013),National

Scholarship Foundation of China, Promotion plan for young teachers' scientific research ability of Minzu University of China, Master's independent research project of Minzu university of China (182164) and the MUC 111 Project are gratefully acknowledged.

References

[1] R. M. Haralick, L. G. Shapiro. Image segmentation techniques [J]. *Computer vision, graphics, and image processing*, 1985, 29(1): 100-132.

[2] J. Shi, J. Malik, Normalized cuts and image segmentation [J]. *IEEE transactions on pattern analysis and machine intelligence*, 2000, 22(8): 888-905.

[3] C. L. Zhao, Image segmentation based on fast normalized cut [J]. *Open cybernetics and systemics journal*, 2015, 9(1):28-31.

[4] Y. Boykov, O. Veksler, R Zabih. Fast approximate energy minimization via graph cuts [J]. *IEEE transactions on pattern analysis and machine intelligence*, 2001, 23(11): 1222-1239.

[5] D. S. Wen. Image semantic segmentation based on weakly supervised learning of two clustering [J]. *Foreign electronic measurement technology*, 2017, 36(9):30-34.

[6] Y. Liu, J. Liu, Z Li, et al. Weakly-supervised dual clustering for image semantic segmentation[C]. *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on. IEEE*, 2013: 2075-2082.

[7] R Guo, X P Yang, J Wang. The optimization analysis of the image segmentation and denoising based on the improved FCM clustering algorithm [J]. *CAAI Transactions on Intelligent Systems*, 2016, 11(6):227-233.

[8] X. X. Zhang. Research on image semantic segmentation based on probability Graph Model [D]. Xiamen University, 2014.

[9] C. F. Zhang. Image semantic segmentation based on conditional random filed [J]. *Computer CD software and applications*, 2012(9):21-23.

[10] X. M Zuo, Z. Zhao, T. T Gou. RGB-D image segmentation method based on interactive conditional random fields [J]. *Computer applications and software*, 2017, 34(3):174-180.

[11] L. Mao, M. Xie. Image semantic segmentation based on higher-order CRF model [J]. *Application research of compute*, 2013, 30(11):3514-3517.

[12] L. J. Wang, Y Q Zhong, H Guo, et al. Improved image segmentation algorithm based on order conditional random field model [J]. *Computer engineering*, 2016, 42(6):241-246.

[13] K Chen. Image semantic segmentation combine bottom segmentation [D]. Shanghai Normal University, 2013.

[14] X. Yang, Y. Fan, L Gao. Image semantic segmentation based on texture element block recognition and merging [J]. *Computer engineering*, 2015, 41(3):253-257.

[15] W. F. Zhang. Semantic segmentation algorithm based on random forests [J]. *School of optical-electrical and computer engineering*, 2017, 30(2):72-75.

[16] P. Cao, J H Qian, Z Chen, X H Li. Image semantic segmentation method based on image hierarchical tree [J]. *Application research of computer*, 2017:3379-3384.

[17] E. Shelhamer, J Long, T Darrell. Fully convolutional networks for semantic segmentation [J]. *IEEE transactions on pattern analysis and machine intelligence*, 2014, 39(4):640-651.

[18] H X Chen. Semantic segmentation based on convolutional neural networks [D]. Zhejiang University, 2016.

[19] H. Noh, S. Hong, B Han. Learning deconvolution network for semantic segmentation[C]. *Proceedings of the IEEE International Conference on Computer Vision*, 2015: 1520-1528.

[20] H. Zhao, J. Shi, X Qi, et al. Pyramid scene parsing network[C]. *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017: 2881-2890.

[21] G. Lin, A. Milan, C. Shen, et al. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation[C]. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017:5168-5177.

[22] S. Hong, H Noh, B Han. Decoupled deep neural network for semi-supervised semantic segmentation[C]. *Advances in neural information processing systems*, 2015: 1495-1503.

[23] V. Badrinarayanan, A Kendall, R Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation [J]. *IEEE transactions on pattern analysis and machine intelligence*, 2017, 39(12): 2481-2495.

[24] S J. gou, M Drozdzal, D Vazquez, et al. The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation[C]. *Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on. IEEE*, 2017: 1175-1183.

[25] A. Paszke, A Chaurasia, S Kim, et al. Enet: A deep neural network architecture for real-time semantic segmentation [J]. arXiv preprint arXiv:1606.02147, 2016.

[26] M A Islam, M Rochan, N D B Bruce, et al. Gated feedback refinement network for dense image labeling[C]. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2017: 4877-4885.

[27] N. Souly, C Spampinato, M Shah. Semi and weakly supervised semantic segmentation using generative adversarial network [J]. arXiv preprint arXiv:1703.09695, 2017.

[28] K. He, G Gkioxari, P Dollor, et al. Mask r-cnn[C]. Computer Vision (ICCV), 2017 IEEE International Conference on. IEEE, 2017: 2980-2988.

[29] Y Li, H Qi, J Dai, et al. Fully convolutional instance-aware semantic segmentation[C]. *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017: 2359-2367.

[30] J. Dai, K He, J Sun. Instance-aware semantic segmentation via multi-task network cascades[C]. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016: 3150-3158.

[31] K. He, X Zhang, S Ren, et al. Deep residual learning for image recognition[C]. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016: 770-778.

[32] O. Ronneberger, P Fischer, T Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation[C]. International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, Cham, 2015:234-241.

[33] D. Liu, X. J Liu, M Z Wang. Semantic Segmentation with Multi-scale Convolutional Neural Network [J]. *Remote Sensing Information*, 2017, 32(1):57-64.

[34] D Martin, C Fowlkes, D Tal, et al. A Database of Human Segmented Natural Images and its Application to Evaluating Segmentation Algorithms and Measuring Ecological Statistics[C]. *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on. IEEE*, 2002:416-423.

[35] S. Song, S. P. Lichtenberg, J Xiao. SUN RGB-D: A RGB-D scene understanding benchmark suite[C]. *Computer Vision and Pattern Recognition*, IEEE, 2015:567-576.

[36] M Everingham, L V Gool, C K I Williams, et al. The Pascal, Visual Object Classes (VOC) Challenge [J]. *International Journal of Computer Vision*, 2010, 88(2):303-338.

[37] M. Cordts, M Omran, S Ramos, et al. The Cityscapes Dataset for Semantic Urban Scene Understanding[C]. *Computer Vision and Pattern Recognition*, IEEE, 2016:3213-3223.

[38] T. Y Lin, M Maire, S Belongie, et al. Microsoft coco: Common objects in context[C]. *European conference on computer vision*, Springer, Cham, 2014: 740-755.

[39] A. Criminisi. Microsoft research cambridge object recognition image database. http://research.microsoft.com/vision/cambridge/recognition,2004.