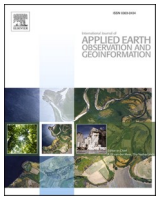
International Journal of Applied Earth Observations and Geoinformation 102 (2021) 102411

Contents lists available at ScienceDirect

International Journal of Applied Earth Observations and Geoinformation

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



A survey on deep learning-based precise boundary recovery of semantic segmentation for images and point clouds

Rui Zhang [a,c], Guangyun Li [b,*], Thomas Wunderlich [c], Li Wang [b]

[a] North China University of Water Resources and Electric Power, 450045 Zhengzhou, China
[b] Department of Geospatial Information, PLA Information Engineering University, 450001 Zhengzhou, China
[c] Chair of Geodesy, Technische Universität München, 80333 München, Germany

ARTICLE INFO

Keywords:
Precise boundary recovery
Semantic segmentation
DCNNs
2D images
3D point clouds

ABSTRACT

Precise localization of semantic segmentation is attracting increasing attention, and salient performances are dominated by deep learning-based methods, especially deep convolutional neural networks (DCNNs). However, the outputs from the final layer of DCNNs are not sufficiently localized for accurate object boundaries due to their invariance properties, which makes precise boundary recovery of semantic segmentation an academically challenging question. Both 2D and 3D objects suffer from the same problem. Considering this, this paper conducts a comprehensive survey of precise boundary recovery for semantic segmentation, focusing mainly on 2D images and 3D point clouds. Firstly, we formulate the problem of potential boundary recovery for semantic segmentation based on DCNNs, elaborate on the terminology as well as background concepts in this field. Then, we categorize boundary recovery methods into four strategies according to their techniques and network architectures to discuss how they obtain accurate boundaries of semantic segmentation. Next, publicly available datasets on which they have been assessed are argued. To compare these datasets, we design diagrams based on five indicators to help researchers judge which are the ones that best suit their tasks. Moreover, we further compare and analyze the performance of all the reviewed methods through experimental results. Finally, current challenges and prospective research issues are discussed extensively.

1. Introduction

Semantic segmentation requires object classification, object detection and boundary localization (Lateef and Ruichek, 2019). It originally applies to 2D images, aiming at a more precise understanding of scenes by assigning a semantic label to each pixel. Since it is defined at the pixel or point level, the assignment of class labels alone is not sufficient; precise localization of each pixel or point is also required. In recent years, with the increased availability and affordability of 3D sensors, including 3D scanners, LiDARs, and RGB-D cameras, 3D data have quickly attracted the increasing interest of researchers. 3D point clouds, as a widely popular 3D data format, can preserve the original geometric information in 3D space without any discretization (Guo et al., 2020). Therefore, it is the preferred representation for many applications related to 3D scene understanding, such as High Definition Mapping (HDM), autonomous driving/drones, and Simultaneous Localization And Mapping (SLAM) (Armeni et al., 2016; Xie et al., 2020). In this regard, semantic segmentation has gradually evolved from being exclusively pixel-wise to including point-wise labeling as well.

Semantic segmentation is, by definition, a dense procedure, hence it requires precise boundary localization of class labels at the pixel-level or point-level. For example, in robot-assisted surgery, pixel-level errors in semantic image segmentation can lead to life-or-death situations (Ulku and Akagunduz, 2019). In autonomous driving, point-level errors in semantic segmentation of point clouds can also cause the same kind of personal injuries. In robotic precision grasping, pixel-level or point-level errors can cause not only grasp failure but even damage to the targets. Therefore, it is extremely crucial for certain applications. This technology is sometimes confused with contour extraction, which belongs to object detection from the perspective of application domains. That is one reason that we review this topic under the title 'precise boundary recovery'.

Both data types (2D images and 3D point clouds) are summarized in this review on precise boundary recovery, and the reasons are as follows. Firstly, 2D images and 3D point clouds complement each other. 3D point clouds can make up for the issues of illumination and posture

* Corresponding author.
E-mail addresses: 599154503@qq.com, guangyun_li_chxy@163.com (G. Li).

https://doi.org/10.1016/j.jag.2021.102411
Received 12 February 2021; Received in revised form 29 May 2021; Accepted 19 June 2021
Available online 18 July 2021

encountered in 2D images and provide rich spatial information for complex scenes as well (Guo et al., 2020), while 2D images can provide extra RGB information. Both of these provide an opportunity for a better and more realistic understanding of the surrounding environments. Second, semantic segmentation of 3D point clouds based on deep learning originates from image-based methods. For example, MVCNN (Su et al., 2015) and VoxNet (Maturana and Scherer, 2015), which transformed 3D point clouds to 2D images and then applied existing knowledge to extract features for point cloud processing. Third, this survey mainly focuses on precise boundary recovery of semantic segmentation. The research objects involve various data formats, including 2D, 2.5D and 3D, while 2D images and 3D point clouds are two of the most important and popular ones for scene understanding in the fields of computer vision, remote sensing, mapping geographic information, navigation and positioning, etc., which have considerable research significance and extensive application prospects. Based on above analyses and the relevance between the two type of data, this survey concentrates on precise boundary recovery of semantic segmentation for 2D images and 3D point clouds.

In recent years, deep neural networks (DNNs) have been proven to excel at a wide range of computer vision and machine learning tasks, e.g. classification, detection, segmentation, etc., among which significant improvements have been achieved by a subset of DNNs known as Convolutional Neural Networks (CNNs). Especially in the past five years, there has been a dramatic increase in global interest in the subject of semantic segmentation (Ulku and Akagunduz, 2019). However, only a few surveys of semantic segmentation using deep learning on 2D images are available, such as (Garcia-Garcia et al., 2018; Lateef and Ruichek, 2019; Ulku and Akagunduz, 2019; Zhao et al., 2017). Additionally, even fewer surveys of deep learning-based semantic segmentation on 3D point clouds have begun to be published in the past two years, such as (Bello et al., 2020; Guo et al., 2020; Xie et al., 2020; Zhang et al., 2019). These surveys mainly focused on semantic segmentation, including background concepts, existing datasets, challenges, description of methods and evaluation of segmentation results to name a few. Regarding accurate boundary recovery for semantic segmentation, Garcia-Garcia et al. (2018) and Ulku and Akagunduz (2019) only briefly mentioned conditional random fields (CRFs) for 2D images. Lateef and Ruichek (2019) simply presented methods using CRF and Markov random field (MRF) and alternative to CRF for 2D images. Additionally, Guo et al. (2020) described an attention mechanism for 3D point clouds in only one paragraph. All of these references are restricted to one aspect and not their key contributions.

To the best of our knowledge, our paper is the first review to focus specifically on deep learning-based precise boundary recovery of semantic segmentation for 2D images and 3D point clouds. Existing research on boundary recovery appears scattered in pieces of literature on semantic segmentation, which makes it very time-consuming and even difficult to keep track of the works. Based on above analyses, this survey is useful either for new researchers who are interested in boundary recovery or for experienced researchers in related fields. It is helpful for new researchers to fully understand the process of development, theories and methods of precise boundary recovery techniques for semantic segmentation, and benchmark datasets on which these methods are assessed. Meanwhile, it is conductive for experienced researchers to obtain related recent advances, grasp the challenges and pay attention to future trends.

Compared with the existing reviews, the main contributions of this survey can be summarized as follows:

(1) A comprehensive review of deep learning-based precise boundary recovery techniques for semantic segmentation for 2D images and 3D point clouds, as shown in Fig. 1.
(2) Fusion of two data types:2D images and 3D point clouds, rather than only one type or other types.
(3) Statistical analysis of benchmark datasets. Histograms, line charts and scatter charts to compare and analyze six public 2D image datasets according to five indicators. Parts are illustrated in Fig. 5 and Fig. 6.
(4) Comparison and analysis between the initial semantic segmentation results and the results after boundary recovery.

The remainder of this paper is organized as follows. Firstly, Section 2 explains the challenges of precise boundary recovery, and clarifies the

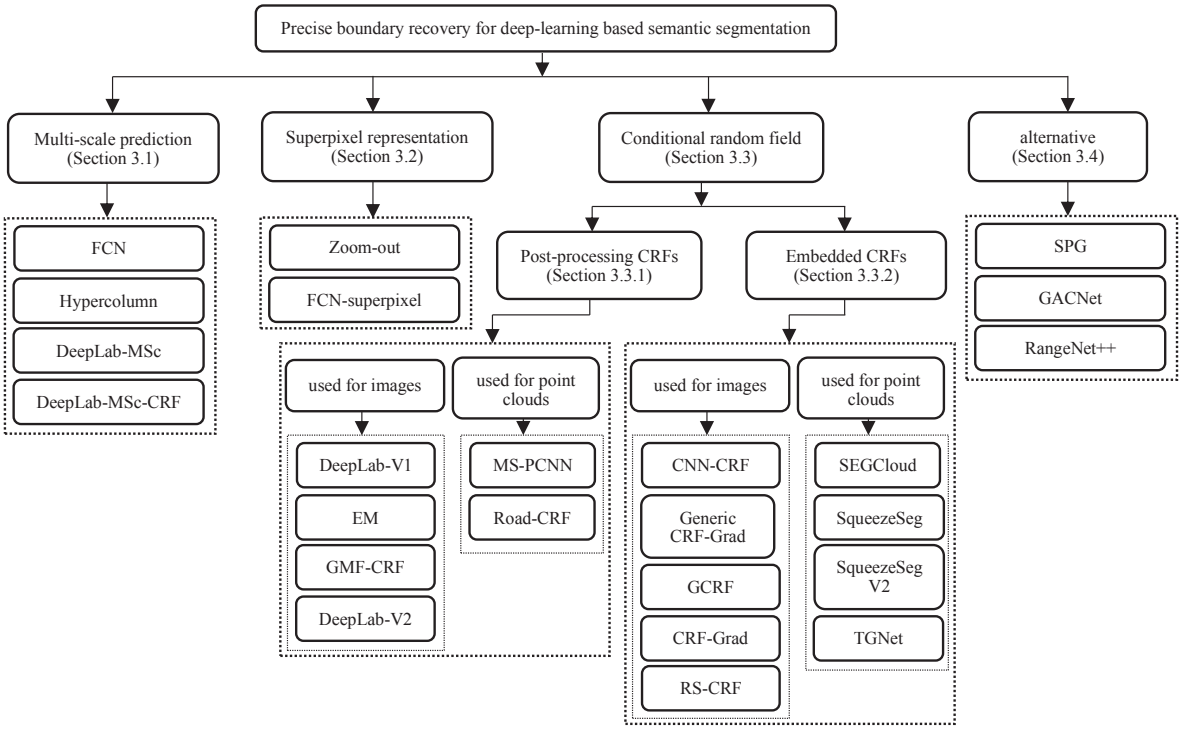


Fig. 1. Visual representation of precise boundary recovery techniques for semantic segmentation.

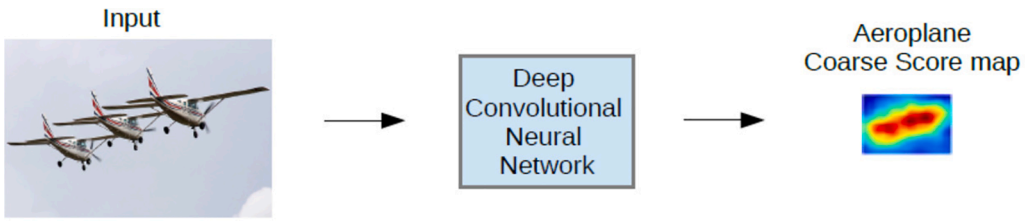


Fig. 2. Coarse score maps of an aeroplane from DCNN before CRF (Chen et al., 2015).

terminology as well as the background concept of semantic segmentation, such as the spatial invariance and smoothing property. Next, Section 3 presents a comprehensive survey of existing precise boundary recovery techniques for semantic segmentation which are grouped into four categories. Then, Section 4 summarizes the benchmark datasets on which the aforementioned methods are assessed. Moreover, it also analyzes the characteristics of benchmark datasets and designs histograms, line charts and scatterplots. Based on these benchmarks, we concentrate on evaluating the performance of deep learning-based boundary recovery models under their experimental results. Following that, challenges and future research directions are discussed. Finally, Section 5 concludes this paper.

2. Terminology and background concepts

To properly understand how precise boundary recovery of semantic segmentation is tackled by DCNNs, it is necessary to become aware of the corresponding terminology and background knowledge. The term boundary in this review is used to describe the object boundary of semantic segmentation. In detail, the boundary of a 2D/3D object refers to the pixels/points located at the outermost part of the object. Likewise, the boundary of a 3D point cloud, for example, represents the collection of points at the outermost part of the point cloud. Boundary recovery of semantic segmentation refers to the boundary optimization techniques used to obtain better segmentation results of boundaries. It is an operation based on the results of preliminary semantic segmentation. The purpose of boundary recovery is to improve the pixel/point segmentation performance of each class, especially for the pixels/points localized on the boundaries. In the field of object segmentation, quantitative relevant literature prefers to describe the improvement of the semantic segmentation of the outermost points as boundary recovery/optimization, as shown in Fig. 3. Therefore, in this review, we keep this terminology in use. This term is also used to distinguish from contour extraction in the field of object detection.

After defining the key terms used in boundary recovery of semantic segmentation, it is then important to ask why the results of semantic segmentation need to be subject to boundary optimization. Semantic segmentation is not an isolated field, but rather a natural step in the progression from coarse to fine inference. The origin could be located at classification, which consists of predicting an input, i.e., predicting the object that appears in an image. Localization is the next step toward fine-grained inference, providing not only the classes but also additional information regarding the spatial location of those classes. Considering this, it is obvious that semantic segmentation is the natural step to achieve fine-grained inference (Garcia-Garcia et al., 2018). However, DCNNs have two properties that are positively detrimental to the inference, one is the spatial invariance and the other is the smoothing property of pooling layers. Spatial information refers to the information having location-based relations with other information. Spatial invariance implies insensitivity to the position of, for example, objects in an image. In the deeper layer of DCNNs, convolutional operations change the representation of the object, so that it is no longer the original. At the last layer, the features extracted by the CNN have no information about their position on the original image. We even lose the information on the pixel size of original objects because of the pooling layers. If we want to get a result of the same size, up-sampling has to be adopted. Up-sampling uses the semantic information of one key pixel to represent the semantic annotation of several pixels surrounding the key one, and the results obtained in this way are inevitably mislabeled, especially for pixels located at the boundaries, which appear jagged. This is the so-called smoothing effect. The goal of semantic segmentation is twofold: classification and precise localization. It is not just classification, nor is it purely a smoothing effect. Therefore, precise boundary recovery of semantic segmentation is the key challenge for improving the accuracy of semantic segmentation, and it is one of the fundamental problems of semantic segmentation based on DCNNs.

For providing a more intuitive view of the effect of precision

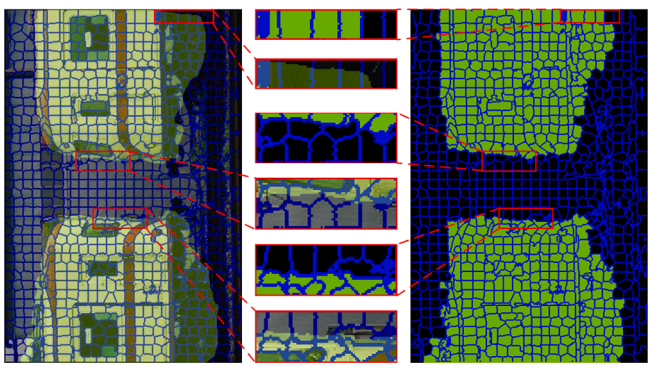


Fig. 3. Boundary optimization and partially enlarged details based on super-pixels (Zhao et al., 2018).

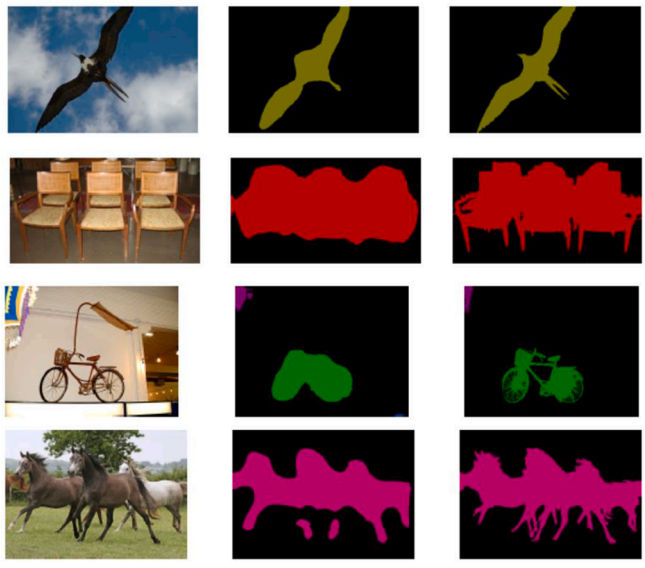


Fig. 4. Visualization results of precise boundary recovery on VOC 2021 *val* set (Chen et al., 2015). For each row, we show the input image, the preliminary segmentation result, and the result after the boundary recovery based on fully connected CRF.

boundary recovery, we show the superpixel-based and CRF-based cases in Figs. 3 and 4, respectively.

3. Precise boundary recovery for semantic segmentation

As illustrated in Fig. 2, DCNN can reliably predict the presence and rough position of objects in an image but are less well suited for pinpointing their exact boundaries of semantic segmentation (Chen et al., 2015). There is a natural trade-off between classification accuracy and localization accuracy with convolutional networks: Deeper models with multiple max-pooling layers have been proven successful in classification tasks, however, their increased invariance and large receptive fields make the problem of inferring position from the scores at their top output levels more challenging.

Based on the algorithm theories and model structures, we classify deep learning-based precise boundary recovery methods of semantic segmentation into four categories. In this section, we will provide a comprehensive review of all four of them at greater length.

3.1. Multi-scale prediction

Multi-scale prediction refers to models in which multiple layers at different scales are concatenated together to improve the boundary localization accuracy of semantic segmentation. For example, FCN, proposed by Long et al. (2015), is a representative multi-scale network, which defined a skip structure that combined coarse, deep layer information with fine, shallow layer information. The skip structure lets the model make more local predictions from shallow layers since their receptive fields are smaller and see fewer pixels. In contemporary works, aiming for precise localization, Hariharan et al. (2015) used hypercolumns as pixel descriptors to capture fine details of the segmentation, which also maintains the high resolution of the lower layers and upsamples the higher layers. DeepLab-V1 (Chen et al., 2015) also explored this multi-scale prediction method, denoted as DeepLab-MSc (Chen et al., 2015), through which the performance on PASCAL VOC 2012 *val* set was improved from 59.8% to 61.3% mean Intersection-over-Union (mIoU). However, it is not as good as DeepLab-MSc-CRF, DeepLab-CRF-LargeFOV and DeepLab-MSc-CRF-LargeFOV (Chen et al., 2015), which adopted post-processing CRFs.

Through the above description, we can see that the multi-scale features can also refine the object boundaries of semantic segmentation. However, the boundary location effect of these methods is only better than the traditional machine learning methods or the plain neural network models with non-multiscale prediction. Subsequent studies usually combine these multi-scale prediction models with superpixels, CRFs, or other boundary refining techniques.

3.2. Superpixel representation

A superpixel can be defined as a group of pixels that share common properties, such as location, color, texture and pixel intensity. Superpixel techniques are widely used for image segmentation due to the following characters: First, superpixels carry more information than pixels. Second, superpixels have a perceptual meaning since pixels belonging to a given superpixel share similar visual properties. Third, their ability to adhere to image boundaries. Based on these characters, superpixel representation is applied to optimize the coarse segmentation boundaries extracted by convolutional neural networks.

Mostajabi et al. (2015) proposed a zoom-out architecture, which utilized simple linear iterative clustering (SLIC) to generate superpixel-level region information, and feature representations were extracted from a sequence of these regions around the superpixels. Then, all the features were combined and fed to a feedforward multilayer network. Although this method is simple, zoom-out lacks the structured nature of the segmentation task. And it is not as effective as CRFs.

Zhao et al. (2018) proposed to optimize the boundaries of semantic segmentation based on superpixels and CRFs. Superpixels were also generated by SLIC. Different from zoom-out method, Zhao et al. (2018) firstly employed the fully convolutional network (FCN) to extract the pixel-level semantic features. After that, the pixel-level information and superpixels were fused to get the boundary-optimized semantic segmentation. Finally, CRF was adopted to get accurate boundaries. To facilitate the comparison and analysis in a later stage, we name it "FCN-superpixel-CRF" in this paper. Besides, we name the combination of FCN-8s with superpixels and CRFs as FCN-superpixel and FCN-CRF, respectively. To evaluate the efficiency of the utility of superpixels and CRFs, Zhao et al. (2018) compared the FCN-superpixel-CRF with the plain FCN-8s model, FCN-superpixel and FCN-CRF, respectively, as shown in Table 1. From the experimental results, we can find that the result of FCN-superpixel is 3.7% better than that of the plain FCN-8s on the PASCAL VOC 2012 *val* dataset. The result with FCN-CRF is 3.6% better than that of FCN-superpixel. And the performance based on superpixels and CRF is further improved by 5% than that of the FCN-CRF, which outperforms the first two boundary recovery techniques by a significant margin. The same trend is seen on Cityscapes. From the description, the combination of superpixels and CRF can achieve the more accurate segmentation result. However, the number of superpixels is set artificially based on prior knowledge, from which the optimal parameter is determined.

From those, we can conclude that the methods based on superpixels have achieved more accurate results comparing with neural network alone. But the number of superpixels is related to the resolution of images. Therefore, this method is not suitable for multi-scale datasets, and parameters affect the accuracy of boundary recovery.

3.3. Conditional random fields

CRFs are the most widely used methods for improving the boundary localization accuracy of semantic segmentation, and CRFs can use the incredible power of CNNs to fine-tune all model features. Additionally, CNNs can more easily capture global properties, such as object shape and contextual information (Kirillov et al., 2015). Therefore, CRFs and CNNs can complement each other's strengths. The combination mode between CRFs and deep convolutional neural networks (DCNNs) can be divided into two subclasses. In one subclass, CRFs are employed as a separate post-processing step disconnected from DCNN training, while in the other, CRFs are formulated as a Recurrent Neural Network (RNN), and thus can be embedded as layers in an existing neural network.

The main advantages of CRFs over other graphical models (such as Markov Random Fields and stochastic grammars) are their conditional nature and their ability to avoid problems of label bias (Lafferty et al., 2001). Consequently, a variety of methods that combine the strengths of CRFs with CNNs to realize accurate localization have been proposed. This section provides comprehensive reviews of the two categories of CRFs (post-processing CRFs and embedded CRFs), and each category is further divided into two subcategories from the perspective of data types: 2D images and 3D point clouds.

3.3.1. Post-processing CRFs

Post-processing CRFs used for Image Semantic Segmentation. Chen et al. (2015) proposed DeepLab-V1, which integrated with fully connected CRFs (FC-CRFs) to improve its localization ability. DeepLab-V1 is a novel direction based on coupling the recognition capacity of DCNNs

Table 1
The mIoU scores of the comparative experiments (Zhao et al., 2018).

Datasets	Plain FCN-8s	FCN-Superpixels	FCN-CRF	FCN-superpixel-CRF
PASCAL VOC 2012	62.2	65.9	69.5	74.5
Cityscapes	56.1	58.9	61.3	65.4

and the fine-grained localization accuracy of FC-CRFs, also known as dense CRFs. Papandreou et al. (2015) developed an alternative method, Expectation-Maximization (EM), for model training under weakly supervised and semi-supervised settings, which decoupled the DCNN and dense CRF training stages and learned the CRF parameters by cross-validation to maximize IoU segmentation accuracy. Although the neural network structures are different, the two models adopted the same fully connected CRF module in the post-processing stage.

Different from CRFs, GMF-CRF (Vemulapalli et al., 2016) utilized a Gaussian CRF model for the task of semantic segmentation by the combination of the Gaussian Mean Field (GMF) and DCNNs. GMF-CRF has the desired property that each of its layers produces an output that is closer to the maximum a posteriori solution of the Gaussian CRF compared to its input. The "patch-patch" model (Lin et al., 2016; Lin et al., 2017) employed the dense CRF as a post-processing method to sharpen the object boundary for generating the final high-resolution prediction. Patch-patch performed approximate training to avoid the repeated inference at every stochastic gradient descent iteration by using piecewise training of CRFs. Unlike other post-processing CRFs, patch-patch learned CNN-based pairwise potential functions for modeling semantic relations between patches. However, the size of the range box was set artificially when constructing the CRF graph, and there was no explicit explanation in the paper as to why it was set to such a size.

DeepLab-V2 (Chen et al., 2017) continuously adopted FC-CRFs and achieved remarkable success in addressing the localization challenge and producing accurate semantic segmentation results. Subsequent researchers pursued efforts in this direction. For example, Wang et al. (2019a) employed the FC-CRF to realize the boundary optimization of buildings and roads on remote sensing images. Zhao et al. (2020) proposed a land cover classification algorithm of polarimetric synthetic aperture radar (SAR) with improved FCN and CRF, where the FC-CRF was used to transfer full image information over global rough classification for fine classification. All these methods employed the FC-CRF as a separate post-processing step disconnected from the DCNN training, where DCNNs first extracted features from input data and then used the outputs as the unary potentials into the fully connected CRF. However, the CRF model considered for precise localization here is a loopy graph, for which the inference is generally computationally expensive due to hundreds of thousands of stochastic gradient descent (SGD) iterations required for training CNNs (Lin et al., 2016).

To further compare and discuss the detailed performance of the above image segmentation models on PASCAL VOC 2012, Table 2 summarizes them based on the following three parameters: GPU type, CRF style, datasets used for evaluation, and performance (mIoU) of before/after CRF. Quantitatively, we can see that DeepLab series localize segment boundaries at a level of accuracy which is beyond previous methods. For DeepLab-V1 (Chen et al., 2015), the mIoU value after CRF is improved to 68.7% from 64.21%, which increases about 4.5% on PASCAL VOC 2012 *val* set, while the value is improved to 71.6% with augmented *trainval* set. The performance of DeepLab-V2 (Chen et al., 2017) is increased by 1.34% after CRF. The difference between DeepLab-V1 and DeepLab-V2 is the backbone network, which changes from VGG-16 to ResNet-101.

Post-processing CRFs used for point cloud Semantic Segmentation. Point cloud semantic segmentation allows finding accurate object boundaries along with their labels in 3D space, which is useful for fine-grained tasks such as object manipulation and detailed scene modeling (Tchapmi et al., 2017). The post-processing CRFs in image boundary recovery of semantic segmentation are gradually extended to 3D point cloud for scene understanding, due to their remarkable success in 2D images. However, the image boundary recovery approaches only consider regular 2D pixel-level data, while 3D point clouds have completely different properties: higher-dimensional, irregular, disordered, unstructured, large-scale and noisy. This means that the precise boundary recovery methods of 2D images cannot be transferred to 3D point clouds directly.

Originally, the combination of CRFs and 3D point clouds is primarily used for the extraction of a single object, rather than the study of boundary optimization in the post-processing inference. 3D CNNs used to process raw point clouds did not appear until 2017, when Qi et al. (2017) designed a novel type of neural network, PointNet, which is a pioneering deep learning framework that can directly consume raw point clouds. From then on, researchers have been beginning to try to combine CRFs with CNNs for boundary refinement for 3D point clouds, and this technique has stepped into a high-speed development stage.

For example, MS-PCNN (Ma et al., 2019) provided an end-to-end feature extraction framework for 3D point cloud segmentation by using dynamic point-wise convolutional operations in multiple scales, then CRF algorithm was developed for improving segmentation boundary accuracy. For CRF inference, it adopted the implementation of (Chen et al., 2017) to 3D point clouds. Li et al. (2020) employed a CNNs-based semantic segmentation method to develop semantic maps of the real-time road scenes by integrating LiDAR and camera information, and then utilized a higher-order 3D CRF model to optimize the semantic map, denoted as Road-CRF. The 3D CRF model defined different smooth terms and added higher-order terms. This method ensures the real-time and accurate requirements, but application scenarios are limited to simpler scenes such as urban roads, and more complex terrain needs to be further considered.

The performance of the methods mentioned above was evaluated with various benchmark datasets, summarized in Table 3. We can see that the styles of CRFs are 3D CRFs, but, none of them is able to consider fine details and long-range contextual information simultaneously.

From the above description, we can see that post-processing CRFs have been a de facto standard in precise boundary recovery for semantic segmentation for a long time (Wang et al., 2019b). However, since CRFs are applied as a separate part following CNNs, the parameters of CRFs and DNNs cannot be optimized simultaneously, resulting in the strengths of both not being exploited. Meanwhile, the above methods concentrate on piecewise training or maximum likelihood learning of restricted model families, such as Gaussian CRFs. For these reasons,

Table 2
Performance of image semantic segmentation models combined with post-processing CRFs on PASCAL VOC 2012. "–" means void, mIoU: mIoU of before/after CRF.

Model	GPU	CRF	mIoU
DeepLab-V1 (Chen et al., 2015)	Titan X	Dense CRF	64.21/68.7 (on *val* set)
			70.3/71.6 (on *val* set, with augmented *trainval* set)
EM (Papandreou et al., 2015)	Tesla k40	Dense CRF	–/71.7 (on *val* set)
GMF-CRF (Vemulapalli et al., 2016)	–	Gaussian CRF	–/73.2 (on *test* set)
patch-patch (Lin et al., 2016; Lin et al., 2017)	–	Dense CRF	–/78.0(on *test* set)
DeepLab-V2 (Chen et al., 2017)	Titan X	Dense CRF	76.35/77.69 (on *val* set)

Table 3
Performance of point cloud segmentation models combined with post-processing CRFs. OA: Overall Accuracy of before/after CRF, mIoU: mIoU of before/after CRF.

Model	GPU	CRF	Datasets	OA	mIoU
MS-PCNN (Ma et al., 2019)	GTX 1080 Ti	3D dense CRF	Paris-Lille-3D	–	–/70.5
			ScanNet	–/87.6	–/56.8
			S3DIS	–/87.3	–/67.8
Road-CRF (Li et al., 2020)	GTX 1070Ti	Higer-order 3D CRF	KITTI	–	–/85.34
			Cityscapes	–	–/73.04

another trend is to explore the combination of the two modeling paradigms, CNNs and CRFs.

3.3.2. Embedded CRFs

In contrast to approaches presented in Section 3.3.1, this section reviews end-to-end frameworks that jointly learn the parameters of CNNs and CRFs.

Transferring CRF as Embedded Layers of CNNs for 2D Images. The idea of formulating the CRF algorithm to a Recurrent Neural Network (RNN) originated from CRF-RNN (Zheng et al., 2015), which was embedded as a part of the CNN to obtain a deep network. The mean-field algorithm it adopted could be traced back to Krahenbuhl and Koltun (2011), depending on which CRF-RNN described it as CNN layers. CRF-RNN used the FCN-8s as its fundamental architecture, which provided unary potentials for the CRF module. The performance was improved by 3.4% mIoU comparing with post-processed CRF on PASCAL VOC 2012 *val* set. From then on, this strategy began to be widely implemented by academia and industry. For example, Motivated by Jancsary et al. (2012) and Tappen et al. (2007),Chandra and Kokkinos (2016) proposed a structured prediction technique that combined the virtues of Gaussian Conditional Random Fields (GCRF) with Deep Learning, which learned features and model parameters simultaneously in an end-to-end FCN training.

Besides, CRF-Grad (Larsson et al., 2017) integrated CNNs with the gradient descent CRF, which was also formulated as RNN layers for scene segmentation. Liu et al. (2019) proposed a network that combined the recognition ability of DCNNs with the fine-grained localization ability of FC-CRFs. We refer to this network as RS-CRF. Here, the coarse segmentation results of the output layer of neural networks were used as input into CRFs to improve the segmentation accuracy of object boundary details in remote sensing images.

The comparison of those methods using CRFs as embedded layers of DCNNs is listed in Table 4 according to three indicators: type of CRF, datasets on which they were assessed, and the performance of these models on evaluated datasets. As we can see, for the type of CRF, in addition to the dense CRF, others adopted the variations of FC-CRFs, such as gaussian CRF and gradient descent CRF. CRF-RNN and GCRF provided both performances of "before-CRF" and "after-CRF". Through the mIoU values and visualization presented in corresponding papers, we can conclude that this end-to-end fashion produces sharp boundaries and dense segmentation. Unfortunately, however, the drawbacks of embedded CRFs need further investigation, such as the multi-scale problem. Although embedded CRFs can integrate contextual information, they do not take into account the size of the objects.

Transferring CRF as Embedded Layers of CNNs for 3D Point Clouds. To extend the strengths of the combination of CNNs and CRFs to the 3D space, many studies have attempted to extend 2D CRF-RNN (Zheng et al., 2015) to higher-order 3D CRFs applicable for point clouds.

For instance, SEGCloud (Tchapmi et al., 2017) combined the advantages of NNs, trilinear interpolation (TI) and the FC-CRF to obtain 3D point-level segmentation based on voxel predictions. Firstly, 3D point clouds were voxelized, and coarse voxel predictions from a 3D NN were transferred back to the raw 3D points via TI, then 3D FC-CRF was used to provide fine-grained labels for 3D points. The purpose of voxelization is to reduce memory consumption and simplify the semantic labeling process because all 3D points within a voxel are assigned the same semantic label. However, voxelization limits the resolution of semantic labels at the CRF stage, and thus the low resolution leads to information loss. It is difficult to keep a balance between memory requirements and an adequate representation of the 3D space without information loss.

Wu et al. (2018) proposed a CNN-based end-to-end pipeline, SqueezeSeg, to address the semantic segmentation of road objects in 3D LiDAR point clouds, including only *car*, *pedestrian* and *cyclist*. The output of the CNN, a point-wise label map, was then refined by a CRF implemented as a recurrent layer. After CRF, the overall segmentation accuracy on the KITTI dataset was improved, with the segmentation performance (IoU) of the *car* category increasing from 60.9% to 64.6%. However, the accuracy of the other two categories, *pedestrian* and *cyclist*, decreased instead. In 2019, the second version, SqueezeSegV2 (Wu et al., 2019), was proposed. It focused on the improvement of network structure, while the CRF layer was removed. Later, Milioto et al. (2019) validated the overall performance of SqueezeSeg and SqueezeSegV2 on the KITTI *test* set, shown in Table 5. The segmentation performance of SqueezeSeg is 29.5% before CRF and 30.8% mIoU after CRF respectively, which are marginally different, and the performance of SqueezeSegV2 has no improvement after CRF. This indicates that not all models will improve their performance after embedded CRFs. Besides, TGNet (Li et al., 2020) proposed a graph convolution architecture to learn expressive and compositional local geometric features from point clouds, which also integrated CRF-RNN (Zheng et al., 2015) for joint training and inference, and achieved 57.8% and 68.17% mIoU on S3DIS and Paris-Lille-3D datasets, respectively. Regrettably, the results before CRF are not provided.

From these descriptions, we can conclude that although integrating CRFs into the original architecture achieves better results in most cases, it is not a substantial improvement. Moreover, embedding CRFs in CNNs is a difficult task because of the additional parameters and high computational complexity required during training.

3.4. Alternatives

With the development of deep neural network architectures, the performance of boundary recovery using postprocessing CRFs and CRF-RNN has been surpassed by some CNNs-based alternatives. For example, Landrieu and Simonovsky (2018) proposed a structure called SuperPoint Graph (SPG) to organize 3D point clouds, which could offer a compact

Table 4
Performance of image semantic segmentation models integrated with embedded CRFs on PASCAL VOC 2012. mIoU: mIoU of before/after CRF.

Model	CRF	mIoU
CRF-RNN (Zheng et al., 2015)	Dense Gaussian CRF	768.3/72.9(on *val* set)
Generic CNN-CRF (Kirillov et al., 2015)	Generic CRF	–/89.01(average accuracy)
GCRF (Chandra and Kokkinos, 2016)	Gaussian CRF	73.86/75.46(on *val* set)
RS-CRF (Liu et al., 2019)	Dense CRF	68.68/-(on *val* set) –/77.2(on *test* set)

Table 5
Performance of point cloud segmentation models integrated with embedded CRFs.OA: Overall Accuracy of before/after CRF, mIoU: mIoU of before/after CRF.

Model	GPU	CRF	Datasets	OA	mIoU
SEGCloud (Tchapmi et al., 2017)	–	3D FC-CRF	S3DIS	–	47.46/48.92 (6-fold cross-validation)
			KITTI	–	35.65/36.78
			Semantic3D	–	58.2/61.3
SqueezeSeg (Wu et al., 2018)	TITAN X	Recurrent CRF	KITTI	–	9.5/30.8 (Milioto et al., 2019)
SqueezeSegV2 (Wu et al., 2019)	TITAN X	Recurrent CRF	KITTI	–	39.7/39.6 (Milioto et al., 2019)
TGNet (Li et al., 2020)	GTX 1080Ti	Recurrent CRF	S3DIS	–/88.5	–/57.8
			Paris-Lille-3D	–/96.97	–/68.17

but still rich representation of contextual relationships between object parts. The results of SPG were then used as the input of the graph convolutional network to extract point cloud features, which achieved 62.1% and 70.8% mIoU on the S3DIS dataset and Semantic3D dataset respectively. SPG described that graph convolution had a similar function with deep learning formulation of CRFs and quantitative experiments and the comparison with CRFs certified this point. Wang et al. (2019b) proposed a graph attention convolution network (GACNet), which shared the same characteristics as CRF by combining the spatial and feature constraints for attentional weights generation. GACNet is equivalent to unfolding the recurrent network of CRF into each layer of the network and also can map the input signals into a hidden feature space for further feature extraction. Experiments verified that GACNet had the same effectiveness as CRF-RNN, but did not show the outstanding advantages. RangeNet++ (Milioto et al., 2019) replaced the CRF with a GPU-based k-Nearest-Neighbor (kNN) search acting directly on the full, unordered point cloud, which enabled the retrieval of labels for all points in the cloud, and achieved accurate boundary recovery and fast semantic segmentation simultaneously. However, this applies to the cases when the samples are evenly distributed. When the samples are unbalanced, the prediction accuracy for rare categories would be low and the retrieval speed would also be very slow.

4. Discussion

4.1. Evaluation of methods

4.1.1. Benchmark datasets and statistical analysis

The availability of public datasets has furthered research on exact boundary recovery for semantic segmentation. For any deep learning-based models and applications, the degree of success is undoubtedly validated by the quality of the datasets used for training. The efficiencies of exact boundary recovery techniques are only comparable and convincing when the models are evaluated with the same benchmarks. For this reason, several datasets assessed using the method presented in Section 3 will be described in further detail.

Representative 2D image benchmark datasets used to evaluate boundary recovery techniques are presented in Table 6. The purpose of this statistical analysis is to provide readers with a deeper understanding of the data architecture and to facilitate the selection of benchmarks for future studies.

In particular, we summarize the point cloud representation, which is one of the core techniques for deep learning-based 3D scene understanding, as shown in Table 7. We find that different datasets adopt different representations, which severely limits their generalization and popularity. If there was a unified standard to represent point cloud features, this would certainly facilitate the rapid development of more advanced technologies based on deep learning and further applications in the industry.

Table 6
2D image semantic segmentation datasets. Classes: semantic classes. Scenes: data acquisition scenes. Resolution: image resolution. Numbers: number of images annotated. For 5,000(20,000), 5,000 is the number of fine labels, and 20,000 is the number of coarse labels.

Dataset	Classes	Scenes	Resolution	Numbers
SIFT-flow (Liu et al., 2009)	33	Outdoor	256*256	2,688
PASCAL VOC 2012 (Everingham et al., 2015)	20	Indoor/ Outdoor	375*500, 500*375	11,530
KITTI (Geiger et al., 2013)	–	Outdoor	1392*512	–
PASCAL-Context (Mottaghi et al., 2014)	459 (59)	Indoor/ Outdoor	Multi-scale	10,103
PASCAL-Part (Chen et al., 2014)	14	Indoor/ Outdoor/ body part	Multi-scale	10,103
MS COCO (Lin et al., 2014)	80	Indoor/ Outdoor	640*512	>200,000
Cityscapes (Cordts et al., 2016)	30	Outdoor	2048*1024	5,000 (20,000)

Datasets are acquired from various scenarios, with different sizes, scales and categories. So, the selection of the right datasets to evaluate and improve our models is crucial. The performance of deep learning models greatly relies on the datasets. Usually, the choice of a dataset is determined by its acquisition environment and application fields. If a model gains superior performance on one dataset, it does not mean that it can achieve the same results on other datasets, because the datasets have different characteristics even if they are labeled as the same class. Therefore, we design a novel approach to evaluate the semantic segmentation models of Section 3 and to compare and analyze the datasets presented in Table 6.

We design histograms, line charts and scatter charts to compare and analyze the six public 2D image datasets listed in Table 6, excluding the KITTI. Firstly, the total number of categories and the total number of instances on the *train* and *val* set of the six datasets are counted. Then we program to calculate the number of categories per image, the number of instances per image, and the correspondence between the number of categories and the number of instances. On this basis, we plot all the statistics, for example, the number of categories per image for six datasets is shown in Fig. 5, while the numbers of images in each of these three datasets for each category are shown in Fig. 6. Our statistic data, statistic codes, diagram codes and all other graphs are publicly available on our project page: https://github.com/zhangrui0828/2D-categoriy-instance-statistics.

Fig. 5 visually illustrates the distribution of categories per image and the distinction among them, with the mean values in parentheses. Based on Fig. 5, we can find that the number of categories in each image on Cityscapes ranges from 4 to 23, and the number of categories in each image on PASCAL-Context ranges from 1 to 24, which means both have a higher complexity than the other four datasets. In contrast to Cityscapes and PASCAL-Context, the maximum number of categories per image on PASCAL VOC 2012 and PASCAL-Part is 6, while only one image contains the maximum category of 18 on the MS COCO dataset, although it has 80 different categories in total.

Next, taking PASCAL VOC 2012 and PASCAL-Context as examples, we further analyze the datasets based on the statistical results. Table 8 presents the experimental results (mIoU) of DeepLab-V2 and CRF-RNN with PASCAL VOC 2012 *val* set and PASCAL-Context dataset. We can see that the semantic segmentation results with PASCAL VOC 2012 are much higher than those with PASCAL-Context, which reflects the influence of the number of categories on semantic segmentation results. More categories mean more complex scenes, and the overall accuracy of semantic segmentation will be lower. For specific segmentation accuracy of each category, the PASCAL VOC 2012 *test* set is taken as an example, as shown in Table 9. Using only the VOC training set, the overall IoU of both DeepLab-V1 and CRF-RNN is above 70%, but the IoU in the category of *chair* is very low, only about 30%. This means that some models, while achieving significant overall performance, still perform poorly in some categories. However, from Fig. 6 we find that the number of images that include the *chair* category is higher than most.

Fig. 6 illustrates the number of images per category in MS COCO, PASCAL-Part and PASCAL VOC 2012. Fig. 6 shows that MS COCO has the most categories of all three datasets and each category appears in much more images. Taking the *person* category as an example, there are 66,808 images in MS COCO, while PASCAL VOC 2012 and PASCAL-Part contain 4,087 and 3,589 images, respectively.

4.1.2. Comparison and analysis of 2D CNN models

The recent state-of-the-art approaches of precise boundary recovery for 2D image semantic segmentation were reviewed in Section 3. Here, we further analyze the efficiency and applicability of these approaches.

Table 7
3D point cloud semantic segmentation datasets. Points: number of points annotated in millions. MLS: Mobile Laser Scanning, TLS: Terrestrial Laser Scanning. Classes: number of labeling. Feature representation: a vector by which each point is represented.

Dataset	Sensors	Ranges	Points	Classes	Scenes	Feature representation
KITTI (Geiger et al., 2013)	MLS	39.2 km	1,799	–	Outdoor	[XYZ, reflectance, label, class]
S3DIS (Armeni et al., 2016)	Structured-light	$6,000m^2$	215	13	Indoor	[XYZ, RGB, Normalized coordinates]
Semantic3D.net (Hackel et al., 2017)	TLS	–	4,009	8	Outdoor	[XYZ, intensity, RGB]
ScanNet (Dai et al., 2017)	RGB-D	34,453 m^2	242	21	Indoor	[XYZ, RGB, label]
Paris-Lille-3D (Roynard et al., 2018)	MLS	2 km	143	50	Outdoor	[XYZ, xyz_origin, GPS_time, reflectance, label, class]

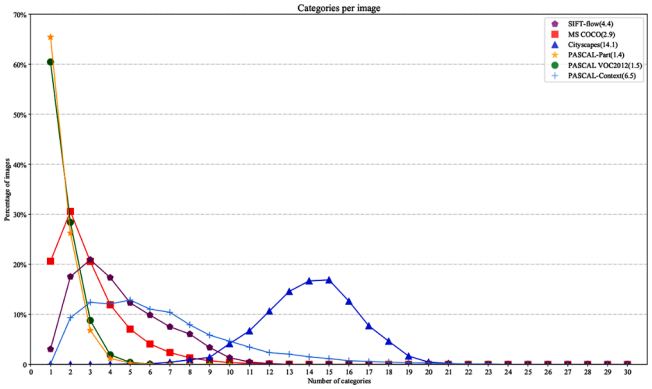


Fig. 5. Categories per image.

Table 10 details the performance descriptions of the methods presented in Section 3, which are tested on PASCAL VOC 2012, PASCAL-Context, PASCAL-Part, Cityscapes and SIFT-flow. We can see that all the models are evaluated with PASCAL VOC 2012, except for CRF-Grad (Larsson et al., 2017). The last two columns of Table 10 show the performance of these approaches on PASCAL VOC 2012, *val* set and *test* set respectively. Values in parentheses present the results which do not use MS COCO data for training. Notably, DeepLab-V2 (Chen et al., 2017), which adopted post-processing CRF to refine boundary segmentation, showed superior performance to other models on PASCAL VOC 2012 *val* set. This framework was evaluated with four distinguishable datasets, including data from indoor datasets, outdoor datasets and a body part dataset.

4.1.3. Comparison and analysis of 3D CNN models

The research on point cloud semantic segmentation based on deep neural networks is still ongoing. New ideas and approaches on the topic of 3D deep learning-based frameworks are being increasingly investigated. Current achievements have led to the improvement of the accuracy of 3D point cloud semantic segmentation (Xie et al., 2020).

As described in Section 3, the use of 3D CNNs to directly process raw point clouds began in 2017. Therefore, the approaches that we have reviewed are all relatively up-to-date, originating from 2017 to 2020. And several of them were just published this year. In this section, we further compare and analyze the performance with alternatives.

Table 11 illustrates the performance evaluation of the methods adopting CRFs or alternatives. As can be seen, although adopting the same CRF style (such as embedded CRFs) and evaluating using the same dataset (such as S3DIS), SEGCloud and TGNet achieved different efficiencies of boundary recovery after CRF. This is not because of CRF, but

Table 8
Semantic segmentation results on PASCAL VOC 2012 *val* set and PASCAL-Context.

Model	PASCAL VOC 2012 *val* set	PASCAL-Context
DeepLab-V2 (Chen et al., 2017)	77.69	45.7
CRF-RNN (Zheng et al., 2015)	72.9	39.28

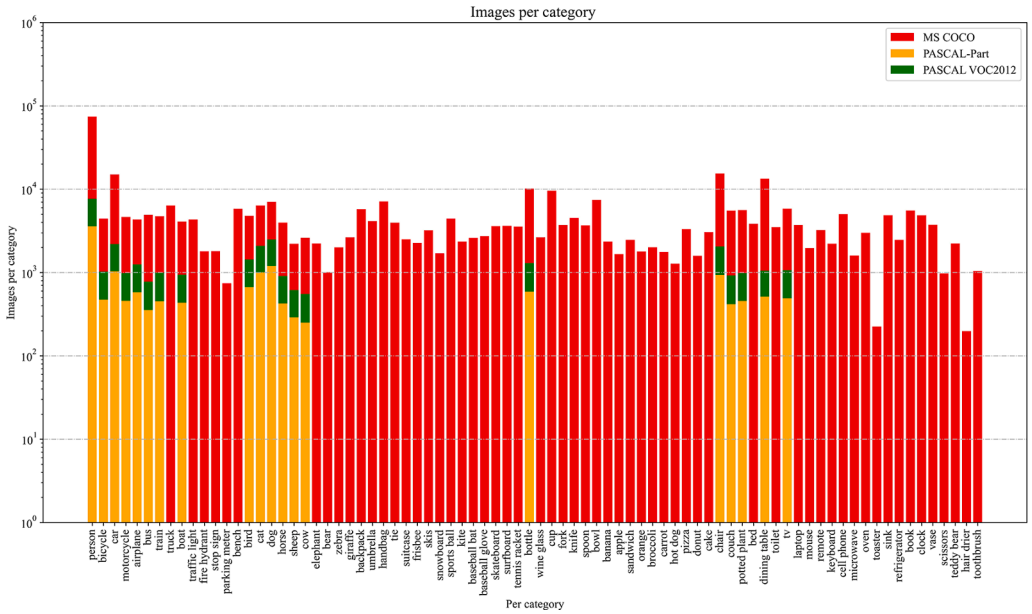


Fig. 6. Number of images per category in MS COCO, PASCAL-Part and PASCAL VOC 2012. The bars are overlaid on the same categories.

Table 9
Individual category results on PASCAL VOC 2012 *test* set.

Model	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mIoU
DeepLab-V1 (Chen et al., 2015)	84.4	54.5	81.5	63.6	65.9	85.1	79.1	83.4	30.7	74.1	59.8	79.0	76.1	83.2	80.8	59.7	82.2	50.4	73.1	63.7	71.6
CRF-RNN (Zheng et al., 2015)	87.5	39.0	79.7	64.2	68.3	87.6	80.8	84.4	30.4	78.2	60.4	80.5	77.8	83.1	80.6	59.5	82.8	47.8	78.3	67.1	72.0

the neural network structures themselves are different, resulting in different semantic segmentation results, i.e., the input to the unitary potential energy of CRF is already different. Taking the subsampling techniques used in the network architecture as an example, SEGCloud used a random subsampling of points in highly dense datasets, while TGNet conducted the farthest point sampling (FPS) algorithm to subsample the point set with a family of ratios. The last three lines illustrate that three alternative methods, including graph convolution, graph attention convolution and kNN approach, also achieve good boundary recovery efficiencies.

Among all the approaches listed in Table 11, five of them had been evaluated with the S3DIS. For fully comparing their performance, the percentage points of mIoU on S3DIS are summarized in the penultimate column. From 2017 till now, significant improvements have been achieved in semantic segmentation based on 3D point clouds, improving from 48.92% to 62.85. However, most of them are still lacking higher representativeness and remarkable robustness. Although CNNs have become the *de facto* standard for semantic segmentation, they have not yet brought a true breakthrough for 3D point clouds. The related researches are still very limited and in the infant stage compared to 2D semantic image segmentation. The main challenge is the thorough and efficient extraction of high-level 3D point cloud features, specifically in large-scale and complex outdoor environments.

4.2. Challenges

Through the above discussion, we find that the following issues need to be further investigated for boundary recovery of semantic segmentation.

1. *Hybrid framework*. For precise boundary recovery of semantic segmentation, hybrid strategies are efficient solutions. One key problem is which modules to choose and how to integrate them. A hybrid framework usually includes at least two parts, one is used to segment coarsely and the other to recover accurate boundaries. A good example of this is the fusion of CNNs and RNNs. For instance, ReSeg (Visin et al., 2016) fed the input image to a VGG-like CNN encoder and then processed afterward by recurrent layers (namely the ReNet architecture) to better localize the pixel labels. Another example is the hybrid dilated convolution (HDC) framework (Wang et al., 2018), which enlarged the receptive fields of the network to aggregate global information in the encoding phase.
2. *Raw point cloud-based boundary recovery*. Boundary recovery techniques originate from 2D image processing and then are transformed and applied to 3D point clouds through Transfer Learning. One of the popular solutions is dimensionality reduction. To be more specific, 3D point clouds are projected into 2D images from multiple perspectives. Then 2D CNNs are adopted to extract features from each view. Finally, 2D semantic segmentation results are projected back to 3D point clouds. Thereby, 3D semantics can be acquired. However, this method would cause numerous limitations and lead to the loss of a large number of important geometric spatial information, which finally affects the accuracy of point cloud segmentation, and it is also seriously influenced by the angle of projection (Zhang et al., 2019). Moreover, in some cases, the algorithms suitable for 2D images cannot be applied to 3D point clouds directly. Consequently, the research directly based on raw point clouds is still in its infancy and has significant potential for development, especially for large-scale, sparse, or unbalanced point clouds.
3. *Criterion for annotating datasets*. There are mainly three methods for the annotation of 2D images and 3D point clouds: (1) manual labeling (Hackel et al., 2017; Roynard et al., 2018), (2) Combining models with human assistance, and (3) crowdsourcing (Dai et al., 2017). For the second method, a segmentation model is first used to obtain the coarse labels, and then the fine labels are obtained with manual assistance. However, the criterion for annotating datasets is

Table 10
Performance comparison of reviewed 2D CNN models with different boundary recovery strategies.

Boundary recovery strategy	Model	PASCAL VOC 2012	PASCAL-Context	PASCAL-Part	Cityscapes	SIFT-flow	mIoU (with *val* set)	mIoU (with *test* set)
multi-scale prediction	FCN-8s (Long et al., 2015)	•					–	(62.2)
	Hypercolumns (Hariharan et al., 2015)	•					–	(62.6)
	DeepLab-MSc (Chen et al., 2015)	•					61.3	–
superpixel representation	Zoom-out (Mostajabi et al., 2015)	•					–	69.6(64.4)
	FCN-superpixel-CRF (Zhao et al., 2018)	•			•		74.5	–
Post-processing CRFs	DeepLab-V1 (Chen et al., 2015)	•					68.7	(71.6)
	EM (Papandreou et al., 2015)	•			•		71.7	–
	GMF-CRF (Vemulapalli et al., 2016)	•					–	73.2
	patch-patch (Lin et al., 2016; Lin et al., 2017)	•	•			•	–	78.0(75.3)
	DeepLab-V2 (Chen et al., 2017)	•	•	•	•		77.69	–
Embedded CRFs	CRF-RNN (Zheng et al., 2015)	•	•				72.9(69.6)	74.7(72.0)
	GCRF (Chandra and Kokkinos, 2016)	•					75.46	–
	CRF-Grad (Larsson et al., 2017)				•		–	–
	RS-CRF (Liu et al., 2019)	•					–	77.2

Table 11
Performance evaluation of the reviewed 3D CNN models. The penultimate column shows the percentage points of mIoU on S3DIS."*" denotes the 6-fold cross validation is used when it is evaluated with S3DIS dataset.

Boundary recovery strategy	Model	S3DIS	Semantic3D.net	Paris-Lille-3D	KITTI	ScanNet	6-fold cross validation	mIoU
Post-processing CRF	MS-PCNN (Ma et al., 2019)	•		•		•		67.8
	Road-CRF (Li et al., 2020)				•			–
Embedded CRFs	SEGCloud (Tchapmi et al., 2017)	•	•		•		*	48.92
	SqueezeSeg (Wu et al., 2018)				•			–
	SqueezeSegV2 (Wu et al., 2019)				•			–
	TGNet (Li et al., 2020)	•		•				57.8
Alternatives	SPG (Landrieu and Simonovsky, 2018)	•					*	62.1
	GACNet (Wang et al., 2019b)	•	•					62.85
	RangeNet++ (Milioto et al., 2019)				•			–

not uniform, resulting in different data representations and file formats, which can be time-consuming in the preprocessing stage. As shown in Table 7, vectors of different dimensions were used for feature representation. It is imperative to develop a unified labeling criterion or industry standard.

4. *Tremendous performance gap among different categories.* The tremendous segmentation performance gap among different categories is still a significant challenge. For example, DeepLab-V1 achieved an overall IoU of 71.6% on PASCAL VOC 2012 *val* set, but a very low IoU of 30.7% on the class of *chair* and meanwhile a very high IoU of 85.1% on the class of *bus*. This is a huge gap of 54.4%, as shown in Table 9. The same is true of the CRF-RNN model. Nevertheless, the number of images that include the *chair* category is much higher than most other categories in the PASCAL VOC2012 benchmark. If we could improve the performance of the least effective categories, then undoubtedly the overall performance will be substantially improved.
5. *Data fusion.* The precise boundary recovery methods of 2D images are more mature and easier for realization than 3D point clouds, and not all current algorithms in computer vision can be used for such remote sensing datasets directly. Moreover, the 3D point cloud datasets reviewed are usually multi-source and multi-modal data. Data fusion has become a mainstream trend in remote sensing. For example, Joint 2D-3D-Semantic data provides a variety of modalities including 2D RGB images, 2.5D depth, 3D point clouds and 3D meshes (Armeni et al., 2017). Patra et al. (2018) fused 2D images and 3D depth data obtained from SLAM for road segmentation. Our previous work (Zhang et al., 2018) fused 2D images and 3D point clouds for semantic segmentation of large-scale outdoor scenes.
6. *Interpretability of deep learning.* Deep learning has notable advantages for large-scale and complex scene semantic segmentation. Nevertheless, poor interpretability is its principal shortcoming. Recently, how each type of layer (e.g., convolution, activation, pooling) works is well known. However, the detailed internal decision-making process is not yet completely understood (Xie et al., 2020). If we could have good interpretability of deep learning and fully describe the rationale, we would be able to build the network structure according to the requirements, without having to fine-tune the hyper-parameters based on prior knowledge blindly. Then, the development of deep learning will have a qualitative leap forward, including the application in precise boundary recovery of semantic segmentation.

This review focuses on boundary recovery techniques of semantic segmentation for natural scene understanding, but of course, some cases are not considered. For example, according to Table 7, we can see that aerial point clouds and photogrammetric point clouds are not included in this review from the sensor perspective. From the perspective of application scenarios, only the boundary recovery of natural scene segmentation closely related to people's life is considered, excluding special scenarios such as mountains, tunnels, railway tracks, etc.

5. Conclusion

Although the current prominence of semantic segmentation is dominated by DCNNs, due to the spatial invariance and smoothing properties of convolutional operation, DCNNs are inevitably unfavorable for the precise localization of semantic segmentation. The parts that are localized incorrectly are usually the pixels or points located at the boundary, so fine-grained boundary recovery becomes a key element affecting overall segmentation accuracy. The purpose of this review is to

assess contemporary deep learning-based boundary recovery techniques for improving the performance of semantic segmentation.

To the best of our knowledge, this is the first review that focuses specifically on deep learning-based precise boundary recovery of semantic segmentation for images and point clouds. This paper provides a comprehensive survey of existing precise boundary recovery techniques of semantic segmentation to stimulate future research. It also includes a performance comparison of these techniques, their merits and demerits, the benchmark datasets used for evaluating their performance, and potential challenges. The techniques are surveyed from two perspectives: model structures and data types.

We firstly divided boundary recovery techniques into four categories (multi-scale prediction, superpixel representation, conditional random fields and alternatives), and provided an overview of each category separately. Regarding the third strategy, we further categorized it into two subclasses based on how CRFs were combined with deep network structures.

Furthermore, we described benchmark datasets on which these models were evaluated, summarized their characteristics, and compared their applications. Moreover, we presented the category and instance statistics of image benchmark datasets and designed histograms, line charts and scatterplots to visualize and analyze them. We believe this is novel in the sense that it provides insight into the advantages and disadvantages of these datasets and gives suggestions to researchers about how to choose them. Regarding the review of 3D point cloud datasets, we presented the point cloud representation of each dataset listed in Table 7 for researchers to gain a clearer conception of how the points were represented.

In the end, all the methods reviewed were further evaluated based on the statistical analyses of the benchmark datasets, and we provided useful insight for challenges in this field. The comparison and analysis of 2D CNNs showed that CRFs are a kind of classical boundary recovery method. However, it is difficult to make further significant breakthroughs in the theory of the algorithm itself, and hybrid methods are a direction to explore. The development of precise boundary recovery techniques based on 3D point clouds is even more promising, some of the alternative techniques are especially showing signs of prosperity. There is an irrefutable need for scientific institutions and industry-leading companies alike to pay attention to where the challenges and future directions for boundary recovery lie. We also aim at closing the gap to help unleash the full potential of deep learning approaches for 3D semantic segmentation.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors would like to thank Wolfgang Wiedemann and the members of the English Writing Center of Technical University of Munich for proofreading this paper, and the editor, associate editor and anonymous reviewers for comments and suggestions.

This study is undertaken with the financial support of the National Natural Science Foundation of China (NSFC) (Grant No. 42071454) and the Science and Technology Research Projects of Science and Technology Department in Henan Province, China (No. 192102210265 and 202102210141). Rui Zhang also acknowledges the financial support provided by China Scholarship Council (CSC).

References

Armeni, I., Sax, S., Zamir, A.R., Savarese, S., 2017. Joint 2d–3d-semantic data for indoor scene understanding (J.a.p.a.).

Armeni, I., Sener, O., Zamir, A.R., Jiang, H., Brilakis, I., Fischer, M., Savarese, S., 2016. 3d semantic parsing of large-scale indoor spaces. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1534–1543.

Bello, S.A., Yu, S., Wang, C., Adam, J.M., Li, J.J.R.S., 2020. deep learning on 3d point clouds. Remote Sens. 12, 1729.

Chandra, S., Kokkinos, I., 2016. Fast, exact and multi-scale inference for semantic image segmentation with deep gaussian crfs. In: European conference on computer vision. Springer, pp. 402–418.

Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L., 2015. Semantic image segmentation with deep convolutional nets and fully connected crfs. In: 3rd International Conference on Learning Representations, ICLR 2015, May 7, 2015–May 9, 2015, . International Conference on Learning Representations, ICLR.

Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L., 2017. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Trans. Pattern Anal. Mach. Intell. 40, 834–848. https://doi.org/10.1109/TPAMI.2017.2699184.

Chen, X., Mottaghi, R., Liu, X., Fidler, S., Urtasun, R., Yuille, A., 2014. Detect what you can: Detecting and representing objects using holistic models and body parts. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1971–1978.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B., 2016. The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3213–3223.

Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M., 2017. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5828–5839.

Everingham, M., Eslami, S.A., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A., 2015. The pascal visual object classes challenge: A retrospective. International journal of computer vision 111 (1), 98–136.

Garcia-Garcia, A., Orts-Escolano, S., Oprea, S., Villena-Martinez, V., Martinez-Gonzalez, P., Garcia-Rodriguez, J., 2018. A survey on deep learning techniques for image and video semantic segmentation. Appl. Soft Comput. 70, 41–65.

Geiger, A., Lenz, P., Stiller, C., Urtasun, R., 2013. Vision meets robotics: The kitti dataset. The International Journal of Robotics Research 32, 1231–1237.

Guo, Y., Wang, H., Hu, Q., Liu, H., Liu, L., Bennamoun, M., 2020. Deep learning for 3d point clouds: A survey. IEEE Trans. Pattern Anal. Mach. Intell.

Hackel, T., Savinov, N., Ladicky, L., Wegner, J.D., Schindler, K., Pollefeys, M., 2017. Semantic3d. net: A new large-scale point cloud classification benchmark.

Hariharan, B., Arbeláez, P., Girshick, R., Malik, J., 2015. Hypercolumns for object segmentation and fine-grained localization, in. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 447–456.

Jancsary, J., Nowozin, S., Sharp, T., Rother, C., 2012. Regression tree fields—an efficient, non-parametric approach to image labeling problems. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 2376–2383.

Kirillov, A., Schlesinger, D., Forkel, W., Zelenin, A., Zheng, S., Torr, P., Rother, C., 2015. Efficient likelihood learning of a generic cnn-crf model for semantic segmentation.

Krahenbuhl, P., Koltun, V., 2011. Efficient inference in fully connected crfs with gaussian edge potentials. In: 25th Annual Conference on Neural Information Processing Systems 2011. Curran Associates Inc.

Lafferty, J., McCallum, A., Pereira, F.C., 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

Landrieu, L., Simonovsky, M., 2018. Large-scale point cloud semantic segmentation with superpoint graphs. In: 31st Meeting of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2018, June 18, 2018–June 22, 2018. IEEE Computer Society. pp. 4558–4567. URL: doi: 10.1109/CVPR.2018.00479.

Larsson, M., Arnab, A., Kahl, F., Zheng, S., Torr, P., 2017. A projected gradient descent method for crf inference allowing end-to-end training of arbitrary pairwise potentials. In: International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition. Springer, pp. 564–579.

Lateef, F., Ruichek, Y., 2019. Survey on semantic segmentation using deep learning techniques. Neurocomputing 338, 321–348.

Li, J., Zhang, X., Li, J., Liu, Y., Wang, J., 2020. Building and optimization of 3d semantic map based on lidar and camera fusion. Neurocomputing 409, 394–407. https://doi.org/10.1016/j.neucom.2020.06.004.

Li, Y., Ma, L., Zhong, Z., Cao, D., Li, J., 2020. Tgnet: Geometric graph cnn on 3-d point cloud segmentation. IEEE Trans. Geosci. Remote Sens. 58, 3588–3600. https://doi.org/10.1109/TGRS.2019.2958517.

Lin, G., Shen, C., Van Den Hengel, A., Reid, I., 2016. Efficient piecewise training of deep structured models for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3194–3203.

Lin, G., Shen, C., Van Den Hengel, A., Reid, I., 2017. Exploring context with deep structured models for semantic segmentation. IEEE transactions on pattern analysis and machine intelligence 40 (6), 1352–1366.

Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C. L., 2014. Microsoft coco: Common objects in context. In: European conference on computer vision. Springer, pp. 740–755.

Liu, C., Yuen, J., Torralba, A., 2009. Nonparametric scene parsing: Label transfer via dense scene alignment. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 1972–1979.

Liu, M., Zhang, C., Zhang, Z., 2019. Multi-scale deep convolutional nets with attention model and conditional random fields for semantic image segmentation. In: SPML '19: Proceedings of the 2019 2nd International Conference on Signal Processing and Machine Learning, pp. 73–78. https://doi.org/10.1145/3372806.3372811.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3431–3440.

Ma, L., Li, Y., Li, J., Tan, W., Yu, Y., Chapman, M.A., 2019. Multi-scale point-wise convolutional neural networks for 3d object segmentation from lidar point clouds in large-scale environments URL: doi: 10.1109/TITS.2019.2961060, doi:10.1109/TITS.2019.2961060.

Maturana, D., Scherer, S., 2015. Voxnet: A 3d convolutional neural network for real-time object recognition. In: 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, pp. 922–928.

Milioto, A., Vizzo, I., Behley, J., Stachniss, C., 2019. Rangenet++: Fast and accurate lidar semantic segmentation. In: 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, pp. 4213–4220.

Mostajabi, M., Yadollahpour, P., Shakhnarovich, G., 2015. Feedforward semantic segmentation with zoom-out features. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, June 7, 2015–June 12, 2015. IEEE Computer Society. pp. 3376–3385. URL: doi: 10.1109/CVPR.2015.7298959, doi:10.1109/CVPR.2015.7298959.

Mottaghi, R., Chen, X., Liu, X., Cho, N.G., Lee, S.W., Fidler, S., Urtasun, R., Yuille, A., 2014. The role of context for object detection and semantic segmentation in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 891–898.

Papandreou, G., Chen, L.C., Murphy, K.P., Yuille, A.L., 2015. Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In: Proceedings of the IEEE international conference on computer vision, pp. 1742–1750.

Patra, S., Maheshwari, P., Yadav, S., Banerjee, S., Arora, C., 2018. A joint 3d–2d based method for free space detection on roads. In: 18th IEEE Winter Conference on Applications of Computer Vision, WACV 2018, March 12, 2018–March 15, 2018. Institute of Electrical and Electronics Engineers Inc. pp. 643–652. URL: doi: 10.1109/WACV.2018.00076, doi:10.1109/WACV.2018.00076.

Qi, C.R., Su, H., Mo, K., Guibas, L.J., 2017. Pointnet: Deep learning on point sets for 3d classification and segmentation. In: 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, July 21, 2017–July 26, 2017. Institute of Electrical and Electronics Engineers Inc. pp. 77–85. URL: doi: 10.1109/CVPR.2017.16, doi: 10.1109/CVPR.2017.16.

Roynard, X., Deschaud, J.E., Goulette, F., 2018. Paris-lille-3d: A large and high-quality ground-truth urban point cloud dataset for automatic segmentation and classification. The International Journal of Robotics Research 37 (6), 545–557.

Su, H., Maji, S., Kalogerakis, E., Learned-Miller, E., 2015. Multi-view convolutional neural networks for 3d shape recognition. In: Proceedings of the IEEE international conference on computer vision, pp. 945–953.

Tappen, M.F., Liu, C., Adelson, E.H., Freeman, W.T., 2007. Learning gaussian conditional random fields for low-level vision. In: 2007 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 1–8.

Tchapmi, L., Choy, C., Armeni, I., Gwak, J., Savarese, S., 2017. Segcloud: Semantic segmentation of 3d point clouds. In: 2017 international conference on 3D vision (3DV). IEEE, pp. 537–547.

Ulku, I., Akagunduz, E., 2019. A survey on deep learning-based architectures for semantic segmentation on 2d images arXiv preprint arXiv:1912.10230.

Vemulapalli, R., Tuzel, O., Liu, M.Y., Chellapa, R., 2016. Gaussian conditional random field network for semantic segmentation, in. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3224–3233.

Visin, F., Ciccone, M., Romero, A., Kastner, K., Cho, K., Bengio, Y., Matteucci, M., Courville, A., 2016. Reseg: A recurrent neural network-based model for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 41–48.

Wang, J., Jiansheng, L., Huachun, Z., Xu, Z., 2019a. Extracting typical elements of remote sensing image based on deeplabv3+ and conditional random fields. Comput. Eng. 1–8.

Wang, L., Huang, Y., Hou, Y., Zhang, S., Shan, J., 2019b. Graph attention convolution for point cloud semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 10296–10305.

Wang, P., Chen, P., Yuan, Y., Liu, D., Huang, Z., Hou, X., Cottrell, G., 2018. Understanding convolution for semantic segmentation. In: 2018 IEEE winter conference on applications of computer vision (WACV). IEEE, pp. 1451–1460.

Wu, B., Wan, A., Yue, X., Keutzer, K., 2018. Squeezeseg: Convolutional neural nets with recurrent crf for real-time road-object segmentation from 3d lidar point cloud. In: 2018 IEEE International Conference on Robotics and Automation. Institute of Electrical and Electronics Engineers Inc., pp. 1887–1893. https://doi.org/10.1109/ICRA.2018.8462926

Wu, B., Zhou, X., Zhao, S., Yue, X., Keutzer, K., 2019. Squeezesegv 2: Improved model structure and unsupervised domain adaptation for road-object segmentation from a lidar point cloud. In: 2019 International Conference on Robotics and Automation (ICRA). IEEE, pp. 4376–4382.

Xie, Y., Tian, J., Zhu, X., 2020. A review of point cloud semantic segmentation. IEEE Geosci. Remote Sens. Mag. https://doi.org/10.1109/MGRS.2019.2937630.

Zhang, J., Zhao, X., Chen, Z., Lu, Z., 2019. A review of deep learning-based semantic segmentation for point cloud. IEEE Access 7, 179118–179133.

Zhang, R., Li, G., Li, M., Wang, L., 2018. Fusion of images and point clouds for the semantic segmentation of large-scale 3d scenes based on deep learning. ISPRS J. Photogramm. Remote Sens. 143, 85–96. https://doi.org/10.1016/j.isprsjprs.2018.04.022.

Zhao, B., Feng, J., Wu, X., Yan, S., 2017. A survey on deep learning-based fine-grained object classification and semantic segmentation. Int. J. Autom. Comput. 14, 119–135.

Zhao, Q., Xie, K., Wang, G., Li, Y., 2020. Land cover classification of polarimetric sar with fully convolution network and conditional random field. Cehui Xuebao/Acta Geodaetica et Cartographica Sinica 49, 65–78. https://doi.org/10.11947/j.AGCS.2020.20190038. URL: https://doi.org/10.11947/j.AGCS.2020.20190038.

Zhao, W., Fu, Y., Wei, X., Wang, H., 2018. An improved image semantic segmentation method based on superpixels and conditional random fields. Applied Sciences 8 (5), 837.

Zheng, S., Jayasumana, S., Romera-Paredes, B., Vineet, V., Su, Z., Du, D., Huang, C., Torr, P.H.S., 2015. Conditional random fields as recurrent neural networks. In: 15th IEEE International Conference on Computer Vision, ICCV 2015, December 11, 2015 - December 18, 2015. Institute of Electrical and Electronics Engineers Inc., pp. 1529–1537:. https://doi.org/10.1109/ICCV.2015.17. URL: https://doi.org/10.1109/ICCV.2015.179.