Journal of Ambient Intelligence and Humanized Computing
https://doi.org/10.1007/s12652-022-03829-6

ORIGINAL RESEARCH

Overview of RGBD semantic segmentation based on deep learning

Hongyan Zhang[1] · Victor S. Sheng[2] · Xuefeng Xi[1] · Zhiming Cui[1] · Huan Rong[3]

Received: 20 April 2021 / Accepted: 10 March 2022

Abstract
Semantic segmentation is one of the basic tasks in computer vision. Its purpose is to achieve pixel-level scene segmentation. With the popularity of depth sensors, combining depth data with RGB images for semantic segmentation can improve the accuracy of semantic segmentation. First, this paper mainly summarizes the fusion of RGB information and depth information and then describes the RGBD semantic segmentation method, evaluation metrics, data set, and comparison of the results on the two mainstream data sets, and then make a prospect of possible future research directions, and finally, a conclusion is made. This part of the work has a certain guiding significance for future research on RGBD semantic segmentation and lays a foundation for later research.

Keywords RGBD semantic segmentation · RGBD data set · Deep learning

1 Introduction

As a basic task in computer vision, semantic segmentation has a wide range of applications on autopilot, robotics, and virtual reality. It mainly analyzes and processes images or video frames for certain objectives and tasks. The purpose of semantic segmentation is to distinguish what objects are in the scene and what categories these objects belong to. Traditional semantic segmentation methods mainly deal with RGB images, and the segmentation effect of similar targets is not ideal. In real life, RGB images are easily affected by light, angles, occlusion, etc., and there are problems such as unbalanced categories and label deviations in the data set. As a result, the effect of only using RGB images for semantic segmentation has a larger error with the Ground Truth of segmentation. It is mentioned in the literature that pillows and quilts are of the same color on a bed and will be labeled as the same category, but in fact, pillows and quilts belong to two categories (Jiao et al. 2019). Recently, some semantic segmentation methods have proposed that when combing the depth data with an RGB image, the final effect will be better than the segmentation without depth data. Such segmentation, combining RGB image with depth data is called RGBD semantic segmentation. In the shallow sense, depth data is understood as the distance between sensor and object.

In recent years, a boom of deep learning has been on the rise, and RGBD semantic segmentation tasks have also been developed (Long et al. 2015). From the Fully Convolutional Networks (FCN) (Long et al. 2015) proposed in 2015, to the SegNet (Badrinarayanan et al. 2017) containing the encoder–decoder structure, along with a series of extended networks based on the encoder–decoder structure that appeared in 2016. Later, in 2017, a network that simply connected the depth estimation task and the semantic segmentation task began to appear (Zhang et al. 2018; Liu et al. 2018a, b). All have gradually accelerated the development of semantic segmentation tasks.

✉ Victor S. Sheng
ssheng@uca.edu

✉ Xuefeng Xi
xfxi@usts.edu.cn

Hongyan Zhang
zhy@post.usts.edu.cn

Zhiming Cui
zmcui@usts.edu.cn

Huan Rong
1227558210@qq.com

[1] School of Electronic and Information Engineering, Suzhou University of Science and Technology, Suzhou 215009, Jiangsu, China

[2] Department of Computer Science, University of Central Arkansas, Conway, AR, USA

[3] School of Artifcial Intelligence, Nanjing University of Information Science and Technology, Nanjing, Jiangsu, China

Published online: 07 April 2022

In the early days, RGBD semantic segmentation feature extraction mainly used artificially designed description features (i.e., directional gradient features SIFT, HOG, Surface Normal and contour shape context, etc.). These features were designed according to the characteristics of the data. These artificially designed descriptive features contained human subjective feelings and poor mobility. Studies have shown that combining RGB images with depth data can bring a better segmentation effect (Jiang et al. 2018; Wang and Neumann 2018; Hazirbas et al. 2016). The popularity of depth sensors enables researchers to combine RGB images with geometric depth information to better solve the problem of semantic segmentation of indoor scenes. How to better combine RGB images with depth information is the key difficulty of RGBD semantic segmentation.

This paper mainly reviews the development of RGB-D semantic segmentation based on deep learning in recent years, classifies and summarizes the proposed methods, makes the key points of various networks clear, and describes the later research of various networks, and lists the relevant RGB data sets and evaluation indicators. In addition, with the help of two commonly used data sets, the final segmentation effect of each network in the data set is listed, and the advantages of the multi-task fusion scheme can be seen through comparison and analysis. The future work point can be located in the multi-task fusion. Figure 1 shows the development roadmap of RGBD semantic segmentation methods in recent years.

The main structure of this paper is as follows. In Sect. 2, we mainly introduce the interaction and fusion method of RGB image and depth information, as well as RGBD semantic segmentation-related algorithms. Next, in Sect. 3, the main innovations involved in RGBD related algorithms are introduced. Then, the relevant data sets and evaluation metrics are summarized in Sect. 4. At the same time, some algorithms are compared, and finally, the whole work is discussed and summarized.

2 RGBD semantic segmentation

2.1 Fusion of RGB image and depth information

RGBD semantic segmentation is based on the interaction and fusion of RGB image and depth information to perform semantic segmentation. How to coordinate the relationship between RGB image and depth information is very important. The original method was simply to connect the RGB image and the depth information at the input, and then gradually fused them after extracting the features, or summing the depth information and RGB information in the process of extracting features. However, the researchers discovered the top–down or bottom–up fusion can better improve the accuracy of RGBD semantic segmentation. In other words, in the development process of RGBD semantic segmentation, the fusion of RGB information and depth information can be divided into early fusion, mid-term fusion, late fusion, and multi-level interactive fusion, as shown in Fig. 2. Here, it can be observed that the input is RGB (the red block) and Depth (the orange block), and the network structure is

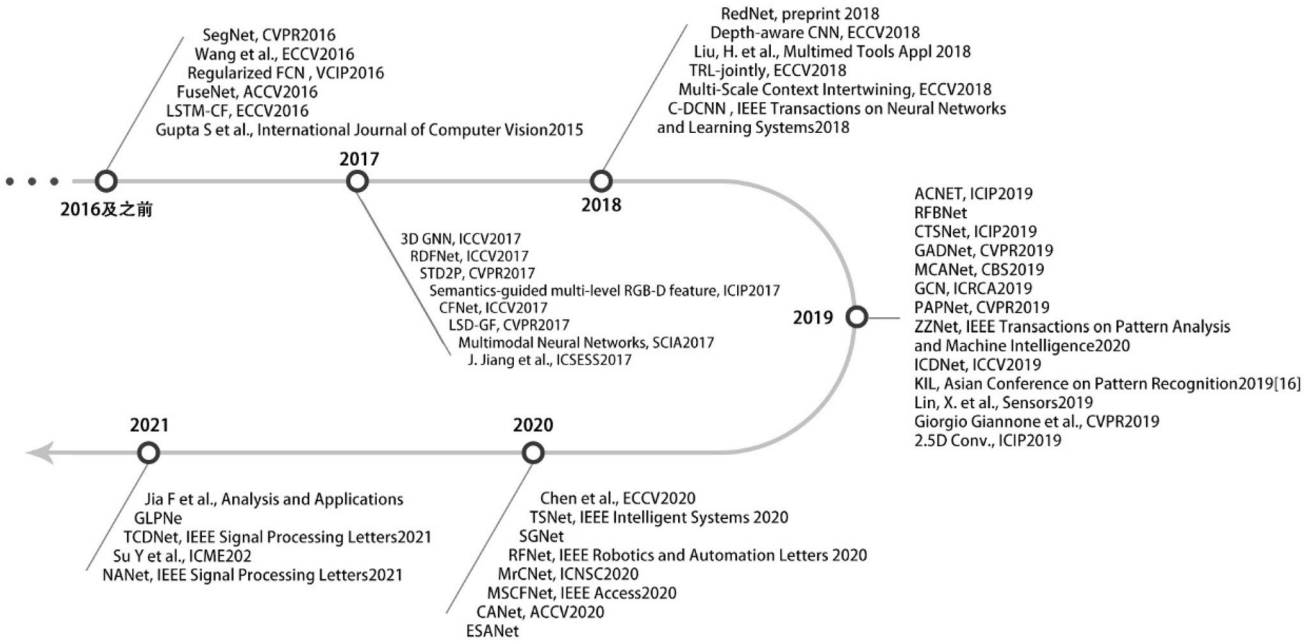

Fig. 1 A road map of RGBD semantic segmentation

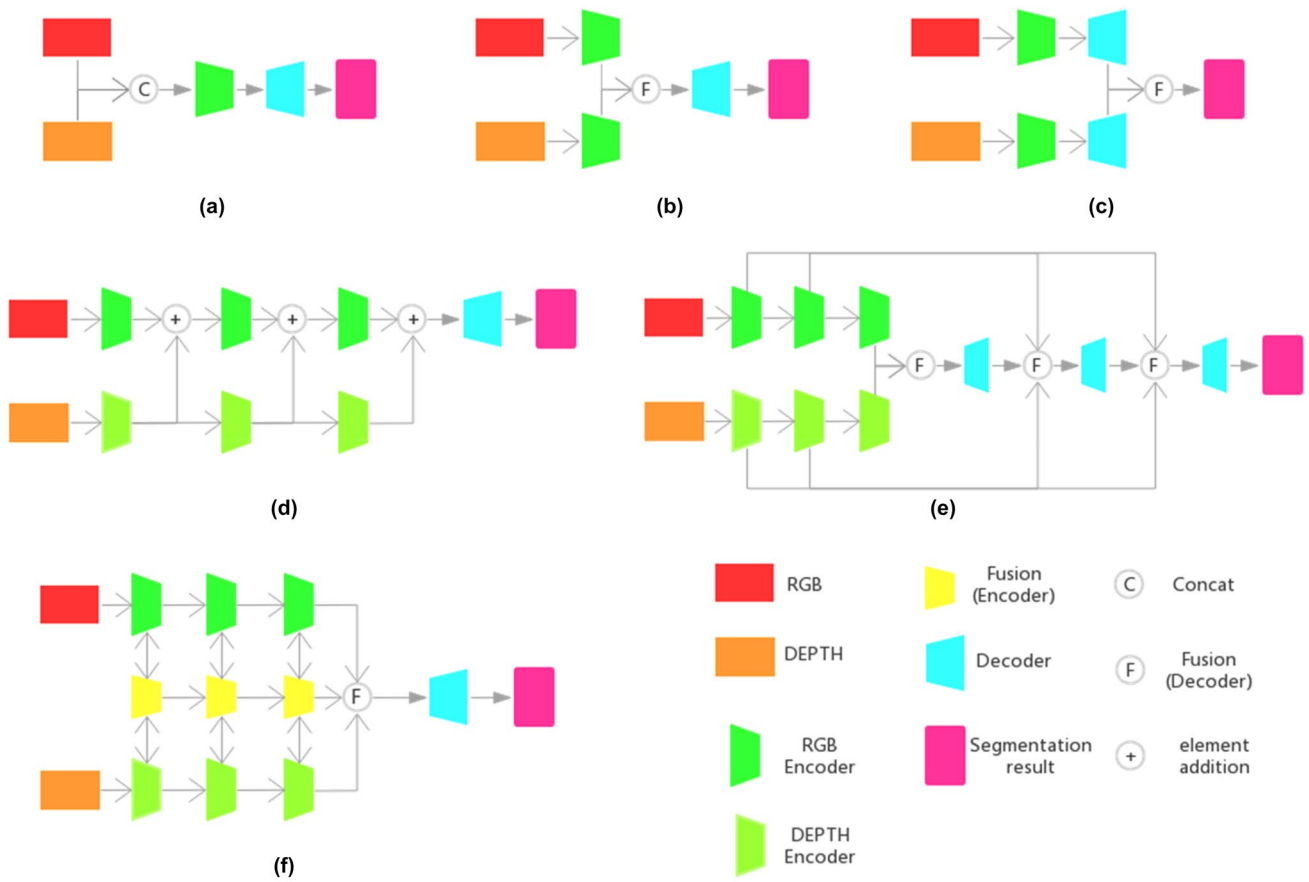

Fig. 2 Fusion of RGB information and depth information (colour figure online)

the Encoder–Decoder (the green block and the blue block) structure. The network usually uses Resnet as the baseline. Next, we will describe one by one from the early fusion.

Early fusion is to fuse RGB information and Depth information before extracting features and then input the results into the convolutional neural network for learning (Couprie et al. 2013), as shown in Fig. 2a. However, such a simple fusion ignores the complementarity between RGB information and depth information and does not make full use of the complementary effect of depth information on RGB images. At the same time, the underlying information in the image is not well corrected.

Mid-term fusion is a concept based on the dual-stream encoder. The RGB branch and the depth branch are respectively extracted by the encoder and then fused. The fusion methods include element summation, concat, and other methods, and then input the fused features into the decoder to obtain the final segmentation result (Cheng et al. 2017), as shown in Fig. 2b. Among them, Wang et al. (2016) proposed to add a feature conversion network between convolution and deconvolution layer, connecting the RGB data with the depth data through the common features between RGB and Depth, as well as their specific features.

Late fusion is to merge the RGB image and the depth image separately after their respective encoding and decoding. They can share the same decoder or use a separate decoder, which has good compatibility. However, but the complementary spatial cues in RGB and depth image advanced features have been weakened after pooling, as shown in Fig. 2c.

Multi-level interactive fusion is relatively more complicated, including multi-level fusion in the feature extraction stage and that in the decoder stage. In the encoder stage, when extracting RGB and depth features separately in the two-stream structure of the network, the depth information is integrated into the RGB branch at multiple levels (Hazirbas et al. 2016; Wang et al. 2020a, b) to better extract RGB information, as shown in Fig. 2d. Although there are certain improvements to semantic segmentation, the interaction between the two-branch encoders is ignored. In addition, another kind of fusion mechanism, or the top-down fusion mode, can also be incorporated into the decoder of the dual-branch structure. In the encoder part, RGB branch and depth branch extract features respectively, and finally fuse them. At the same time, multi-level and multi-scale RGB information and depth information are fused in the decoder (Park

et al. 2017; Wang et al. 2020a, b), as shown in Fig. 2e. In the encoder stage, a three-branch structure of bottom-up fusion also appeared: RGB branch, depth branch, and fusion branch (Deng et al. 2019), as shown in Fig. 2f. There is a two-way interactive fusion between encoders, which makes better use of the dependence and complementarity between branches.

The fusion mechanism of RGB and depth information is diverse and becoming more complex, but at the same time, the segmentation accuracy will gradually improve. The fusion mechanism mainly used is a multi-level interactive fusion mode. These methods are well mixed, and some other elements are added to make the network more accurate and efficient, and at the same time, it is more proper to be adopted for real applications.

2.2 Structure of RGBD semantic segmentation

In this section, we mainly introduce the RGBD semantic segmentation method based on deep learning, mainly described from the following aspects: based on FCN structure, based on Encoder–Decoder structure, multi-task combination. Figure 3 summarizes the approach in this section.

2.2.1 Based on FCN structure

The current mainstream framework is based on CNN. The first proposal is to use CNN-based networks to achieve pixel-level semantic segmentation in Fully Convolutional Networks (Long et al. 2015). FCN directly inputs depth data into the existing RGB semantic segmentation framework and replaces the final fully-connected layer with a convolutional layer. The network can adapt well to any size of the input. At the same time, it can fuse multi-level information with the skip connection to ensure robustness. As shown in Fig. 4.

FCN uses the HHA (Song et al. 2015) proposed by Gupta et al. (2014) to encode depth information at the input. HHA includes horizontal disparity, ground height, and normal angle. Based on FCN, Su et al. (2016) added regularization based on the fully convolutional network to further extend the FCN, and its input is the depth mapping and handcrafted features. By strictly regularizing in fully connected layers, the relationship between these features and their labels can be learned.

Wang et al. (2016) proposed a multi-modal deconvolution structure following the principle that the deconvolution network predicts pixel-level semantics. Wang et al. (2016) used a feature conversion network to associate the common features of the RGB mode and the depth mode, and use their specific features to characterize each. Enhancing public features can better distinguish unique features and enhance robustness. However, this method needs to perform effective noise processing on the mixed features, otherwise, its segmentation effect will be greatly reduced.

The LSD-GF model proposed by Cheng et al. (2017) is mainly composed of a front-end full convolutional network (FCN), an intermediate position-sensitive deconvolutional network, and a final gated fusion layer. The method can expand the receptive field of the network while recovering finer object edge information. In the solution pooling and average pooling operations, an affinity matrix embedded with the paired relationship between adjacent RGB-D pixels is added to restore the clear boundary of FCN mapping. These networks based on FCN expansion have achieved certain results, but at the same time, the high-resolution

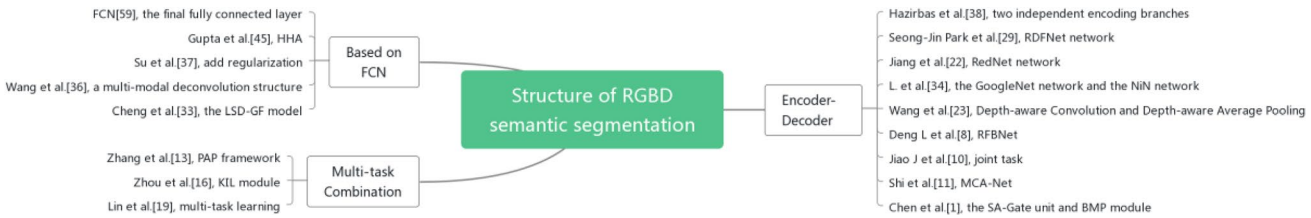

Fig. 3 Summary of deep learning RGBD semantic segmentation methods

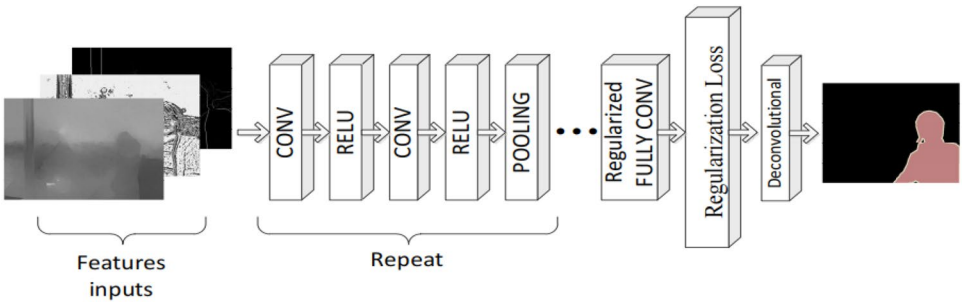

Fig. 4 Overview of regularized FCN framework (Su and Wang 2016)

feature maps were ignored, resulting in the loss of the edge information, and the memory consumption will increase. To enlarge the receiving field and reduce the memory as well as the computing consumption, the Encoder–Decoder structure (Badrinarayanan et al. 2017) followed.

After a given input image, the encoder performs convolution and maximum pooling through a neural network to learn the feature map of the input image. The decoder performs up-sampling and convolution after the encoder has provided the feature map, and gradually realized the category labeling of each pixel.

2.2.2 Based on the Encoder–Decoder structure

Hazirbas et al. (2016) used two independent encoding branches to extract features of depth information as well as RGB information, and continuously integrated the extracted depth features into the RGB feature map as the network deepened. The RDFNet network (2017) adopts the idea of residual learning, and uses the MMFNet (Multi-Modal Feature Fusion Network) block to fuse multi-modal features, and refine the fusion features layer by layer through the RefineNet block. The RedNet network proposed by Jiang et al. (2018) also integrates the residual module into the Encoder–Decoder structure, and adds a pyramid supervised training scheme to alleviate the gradient disappearance problem. As shown in Fig. 5. The Multi-modal NN network proposed by Schneider et al. (2017) uses the GoogleNet network to extract RGB features, and uses the NiN network to extract deep features. After mid-term fusion, the fusion result will be input to the decoder to obtain the segmentation results.

Hazirbas et al. (2016) used two independent encoding branches to extract features of depth information as well as RGB information, and continuously integrated the extracted depth features into the RGB feature map as the network deepened. The RDFNet network (Park et al. 2017) adopts the idea of residual learning, and uses the MMFNet (Multi-Modal Feature Fusion Network) block to fuse multi-modal features, and refine the fusion features layer by layer through the RefineNet block. The RedNet network proposed by Jiang et al. (2018) also integrates the residual module into the Encoder-Decoder structure, and adds a pyramid supervised training scheme to alleviate the gradient disappearance problem. The Multi-modal NN network proposed by Schneider et al. (2017) uses the GoogleNet network to extract RGB features, and uses the NiN network to extract deep features. After mid-term fusion, the fusion result will be input to the decoder to obtain the segmentation results.

Wang et al. (2018) considered that pixels with the same semantic label and similar depth should have a more positive effect on each other, and proposed Depth-aware Convolution and Depth-aware Average Pooling. Depth-aware Convolution uses pixels with a depth similar to that of the central pixel of the kernel to maximize its contribution to the output. Depth-aware Average Pooling allows visual information and geometric information to be transmitted together, and can seamlessly integrate the geometric information reflected by the depth data into the CNN without increasing the number of parameters and calculations.

RFBNet (Deng et al. 2019) proposed a bottom-up interactive fusion structure to model the interdependence between encoders. This method not only gradually aggregates the characteristics of specific modes from the encoder but also calculates complementary characteristics. At the same time, a residual fusion block (RFB) is proposed to represent the interdependence between encoders. The RFB is composed of two specific mode residual units (RU) and a gated fusion unit (GFU). It learns complementary characteristics for specific

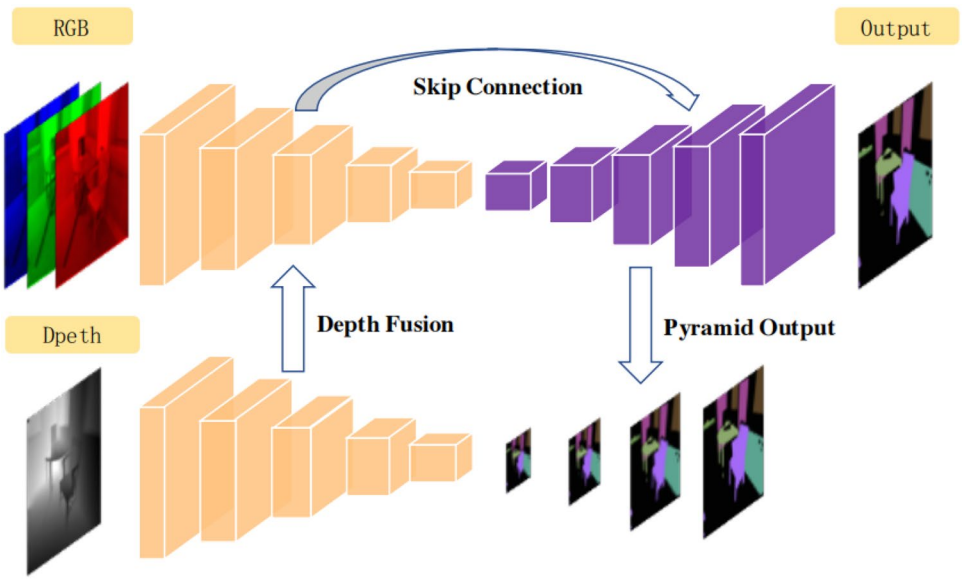

Fig. 5 Overall structure of the RedNet (Jiang et al. 2018)

modal encoders and extracts specific modal characteristics and cross-modal characteristics.

Jiao et al. (2019) decoupled the single-task prediction network into two joint tasks of semantic segmentation and geometric embedding learning. This method extracts geometric perceptual embedding to jointly infer semantic and depth information. The depth map is inferred from the RGB image, and then the inferred depth map is combined with the features extracted from the RGB image to perform the semantic segmentation task. The key idea of the method proposed by Jiao et al. (2019) is to predict the semantic label from an RGB image, while implicitly considering the three-dimensional geometric information based on the RGB image.

Shi et al. (2019) proposed a multi-layer cross-sensing network (MCA-Net) for joint reasoning of two-dimensional appearance and depth geometric information. This method uses the basic residual structure to separately encode texture information and depth geometric information. In addition, a multi-layer cross-sensing fusion module is designed to fuse the multi-scale complementary features extracted from RGB images and depth images. A simple depth conversion method is used to reverse the depth conversion, which overcomes the information hiding problem caused by the distribution of the original depth data.

The key idea of Chen et al. (2020) is to first suppress the features of low-quality depth information, and then use the suppressed features to refine the RGB features. Given RGB-D data as input, the encoder recalibrates and fuses the complementary information of the two modes of RGB and depth through the SA-Gate unit. Then propagates the fused multi-modal features and model-specific features through the bidirectional multi-step propagation (BMP) module, encouraging RGB stream and depth stream to better maintain their specificity during the information interaction process at the encoding stage. Then decode the information through the segmentation decoder network to generate a segmentation map.

These methods mostly combine attention mechanisms and residual learning ideas. When combining high-level features with low-level features, they also pay attention to retaining more edge detail features. However, for the collection of depth information, the result of the collection is still regarded as an ideal state, which is far from the actual situation.

2.2.3 Multi-task combination

The multi-task combination is to decompose the RGBD semantic segmentation task into multiple independent tasks. That is to say, RGBD semantic segmentation can be considered as a joint task of semantic segmentation and depth estimation. This approach is more effective for the extraction of depth information.

Zhang et al. (2019) proposed a pattern affinity propagation (PAP) framework to jointly predict depth, surface normals, and semantic segmentation. The propagation framework performs two types of propagation: cross-task propagation and task-specific propagation, to adaptively propagate those similar patterns. Cross-task propagation integrates a cross-task association mode, which makes the framework adapt to each task by calculating non-local relationships. Next, task-specific propagation performs iterative diffusion in the feature space so that cross-task correlation patterns can be widely spread across tasks. Therefore, the learning of each task can be standardized and promoted through the complementarity of task levels.

Zhou et al. (2019) proposed the Key Knowledge Interactive Learning (KIL) module, which can effectively mine and utilize the connections and complements between semantic segmentation and depth estimation tasks. The encoder consists of a regular layer and some remaining blocks. The decoder has a two-stream network that handles dual tasks. The KIL module can adaptively find the relationship between tasks and interactive information.

Lin et al. (2019) used a single RGB input image to jointly solve the problems of depth estimation and semantic segmentation by a unified convolutional neural network. As shoen in Fig. 6. The method analyzes two different architectures to evaluate which features are more relevant when two tasks

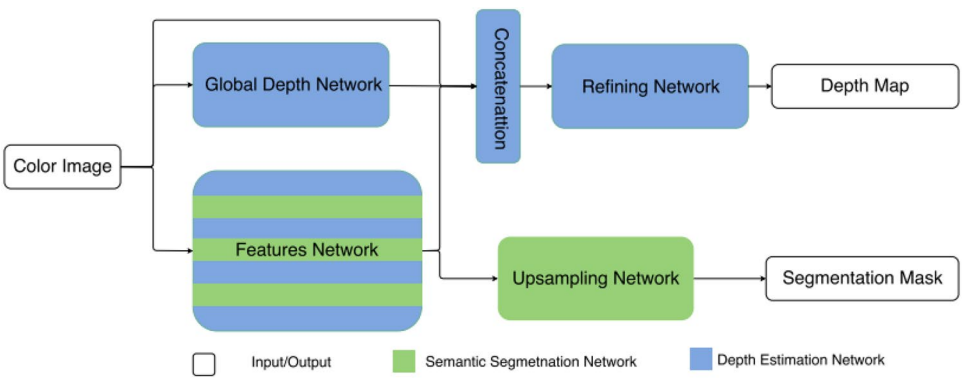

Fig. 6 Multi-task combination network (Lin et al. 2019)

are shared, and which features should be separated to achieve mutual improvement. Finally, they find that solving correlated tasks such as semantic segmentation and depth estimation together can improve the performance of methods tackling the tasks separately. In addition, it is necessary to find a more suitable loss function for a multi-task learning scheme.

It can be understood that the combination of multiple tasks has a certain positive effect on the final result of RGBD semantic segmentation (Wang et al. 2020a, b; Zhang et al. 2019; Zhou et al. 2019; Lin et al. 2019; Liu et al. 2018a; Liu et al. 2018a, b). But at the same time, due to the joint execution of multiple tasks, it will increase the burden on the processor. The memory footprint is too large and the calculation speed is slow. These are the points that need to be further improved in the future.

3 Key technology

This part mainly describes some key points that appear in the improved RGBD semantic segmentation, including attention mechanism, skip connection, spatial pyramid pooling, knowledge distillation, and the application of activation function and loss function. Figure 7 summarizes this section.

3.1 Attention mechanism

The attention mechanism imitates the human visual mechanism. Human vision will devote more attention to a target area while suppressing other temporarily useless information. The attention mechanism in deep learning is similar to the human visual attention mechanism (Kosiorek 2017). The ultimate goal is to find out the more critical information for the current task among the many information.

$$Attention(Q, K, V) = soft\max\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

Note that the mechanism can be understood as a weighted average according to the relational matrix. QK^T is the relational matrix in Eq. 1, and softmax operation is to normalize the relational matrix to the probability distribution, and V is resampled according to the probability distribution, and finally get the new attention result. Attention mechanisms can be divided into soft attention and hard attention based on type. Hard attention is a discrete binary representation that distinguishes which parts are paid attention to and which parts are not paid attention to, focuses more, and emphasizes dynamic change. Soft attention is a continuous representation in the range of 0–1. Continuous values are used to represent the attention degree of each region, and more attention is paid to regions or channels. Soft attention can be divided into spatial domain attention, channel domain attention, and mixed domain attention from dimension. Spatial domain attention transforms the spatial information in the original picture to another space through the spatial conversion module and retains the key information. Channel attention aims to explicitly model the correlation between different channels, automatically obtain the importance of each feature channel through network learning, and finally assign different weight coefficients to each channel, so as to strengthen the important features and suppress the non-important features. Mixed domain attention mixes the spatial domain with the channel domain. RGBD semantic segmentation mostly uses a soft attention mechanism.

At present, the attention mechanism is mostly attached to the Encoder–Decoder structure. In the RGBD semantic segmentation task, some networks use the attention mechanism to improve the semantic segmentation effect. For example, in the three-stream self-attention network (TSNet, three-stream) (Zhou et al. 2020a, b), the network is composed of two asymmetric input streams and a cross-modal distillation stream. As shown in Fig. 8. The cross-modal distillation stream with the self-attention module is used to fuse and refine the characteristics of the RGB stream and depth stream.

In addition, the Attention complementary network (ACNet) (Hu et al. 2019) selectively collects features from the RGB branch and the depth branch. The main contribution is the Attention Complementary Module (ACM) and the

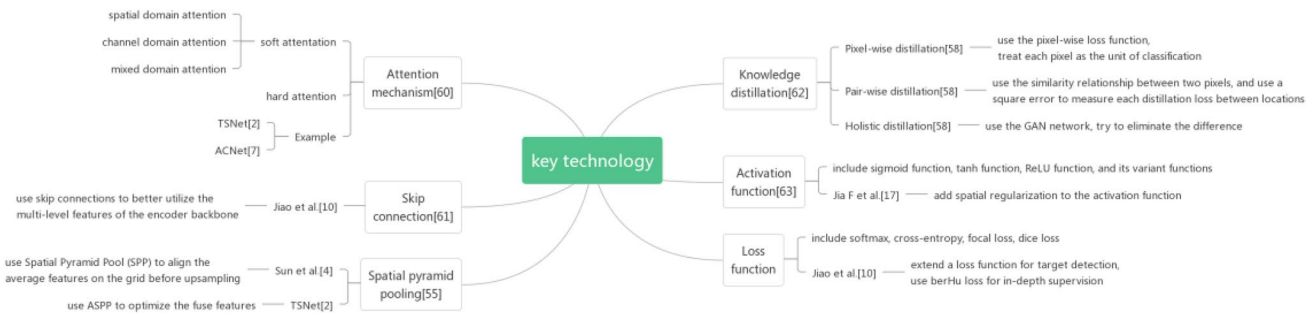

Fig. 7 Summary of the key technology

architecture with three parallel branches. ACM is a channel-based attention module that extracts weighted features from the RGB and depth branches. As shown in Fig. 9.

In a word, the rational use of attention mechanisms can better improve the accuracy of semantic segmentation. At the same time, the emergence of more attention mechanisms will further promote the development of semantic segmentation.

Figure 10 shows the semantic segmentation effect of TSNet (Zhou et al. 2020a, b) with the addition of attention mechanisms and ASPP.

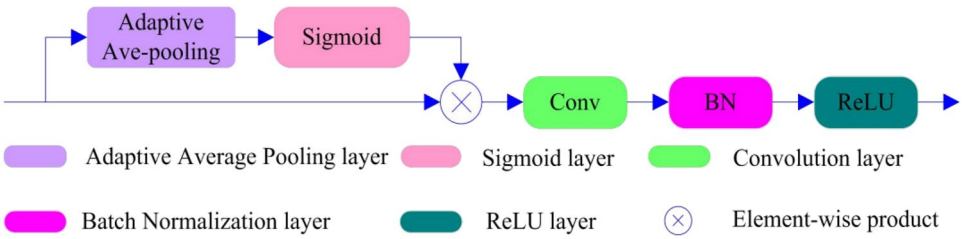

Fig. 8 Structure of the self-attention module[from TSNet (Zhou et al. 2020a, b)]

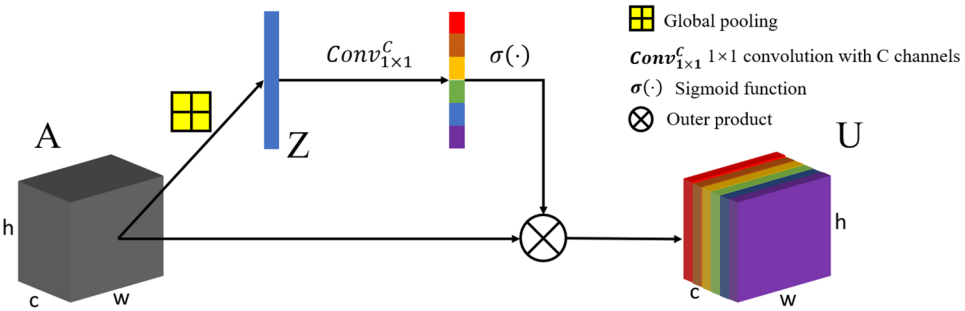

Fig. 9 Attention Complementary Module (ACM) (Hu et al. 2019)

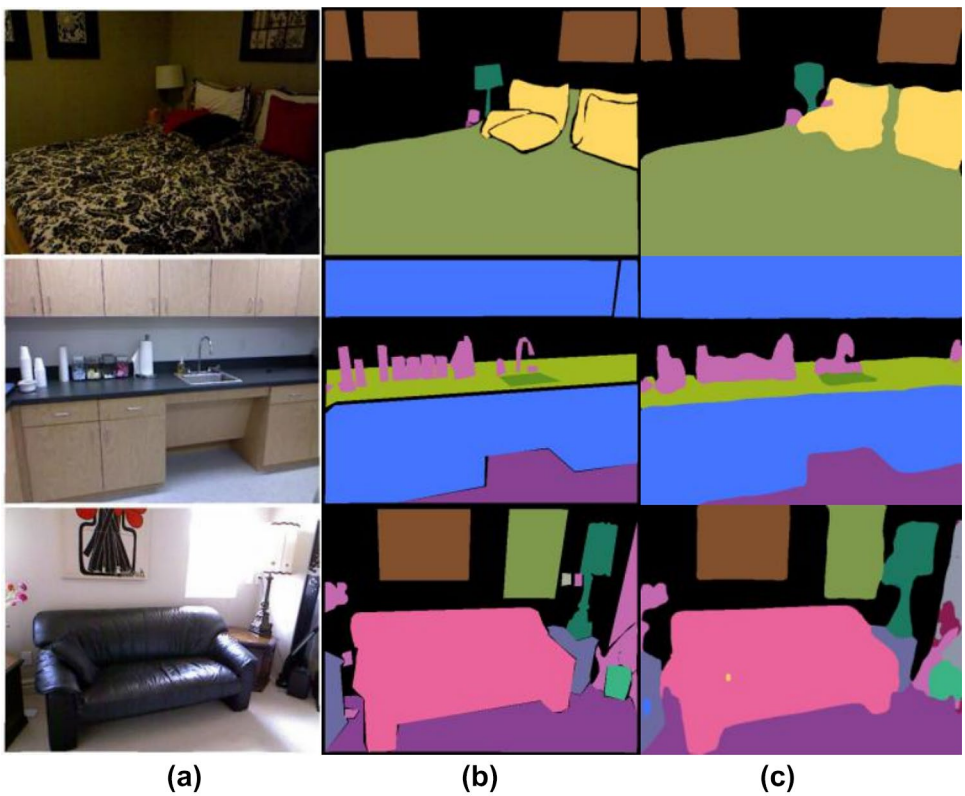

Fig. 10 Results of the proposed TSNet, **a** RGB image, **b** Ground truth, **c** Proposed model[from TSNet (Zhou et al. 2020a, b)]

3.2 Spatial pyramid pooling

Spatial pyramid pooling(SPP) (He et al. 2015) is generally a network layer between the convolutional layer and the fully connected layer, so that feature maps of any size can be converted into fixed-size feature vectors, and different rates can be applied to different layers in parallel. SPP mainly divides the feature map of any size into 16, 4, and 1 blocks at first, and then performs maximum pooling on each block. The pooled features are splice to obtain an output of a fixed dimension to meet the needs of subsequent decoders. Not only does it improve the scale-invariance of images, but it also reduces over-fitting. As shown in Fig. 10.

Sun et al. (2020) used Spatial Pyramid Pool (SPP) to align the average features on the grid before upsampling. The three-stream self-attention network TSNet (Zhou et al. 2020a, b) used ASPP to optimize the fuse features. As shown in Fig. 11.

Pyramid pooling can better extract depth and RGB information, and make a better fusion to improve the accuracy of semantic segmentation.

3.3 Skip connection

Skip connection (He et al. 2016) is often used in residual networks. In a relatively deep network, the skip connection enables the network to fuse the feature map of the corresponding position of the encoder on the channel during the up-sampling process of each level, it can solve the problem of gradient explosion and gradient disappearance during training.

Through the fusion of low-level features and high-level features, the network can retain more high-resolution detailed information contained in high-level feature maps, thereby improving segmentation accuracy. Jiao et al. (2019) used skip connections to better utilize the multi-level features of the encoder backbone so that more details can be enriched and restored in the final semantic feature mapping. Figure 12 shows the semantic segmentation effect before and after the addition of skip connection.

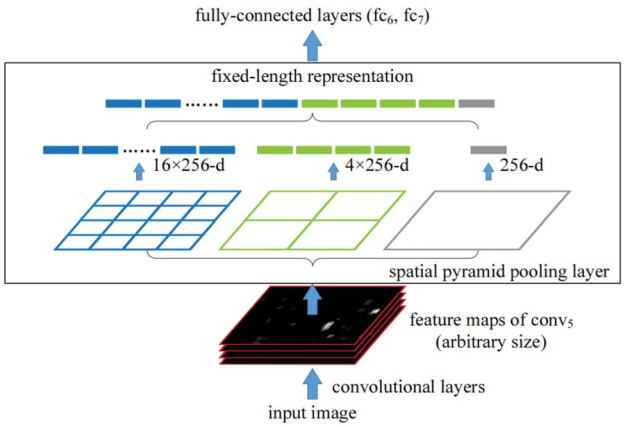

Fig. 11 A network structure with a spatial pyramid pooling layer. Here 256 is the filter number of the conv5 layer, and conv5 is the last convolutional layer [from SPP (He et al. 2015)]

3.4 Knowledge distillation

Knowledge distillation is a method to transform knowledge from a cumbersome model to a compact model to improve the performance of the compact network (Hinton et al. 2015). The key is how to measure the consistency of the output results of the Teacher network and the Student network.

$$q_i = \frac{\exp\left(z_i/T\right)}{\sum_j \exp\left(z_j/T\right)} \tag{2}$$

Equation 2 is the Softmax function with parameter T, where T represents temperature. When the temperature T approaches 0, softmax output converges to a one-Hot vector. When the temperature T tends to infinity, the smoother the output probability distribution of Softmax is, the greater the entropy of its distribution is. The information carried by the negative label will be relatively amplified, and the model training will pay more attention to the negative label. The general method is to train the Teacher network first, and then distill the knowledge of the Teacher network into the Student network at high-temperature T. The objective function in the process of high-temperature distillation was weighted by Distill loss and Student Loss. Adding student Loss can effectively reduce the possibility of being biased by the occasional errors in the teacher network.

The knowledge distillation includes Pixel-wise distillation, Pair-wise distillation, and Holistic distillation (Liu et al. 2019). Pixel distillation uses the pixel-wise loss function to measure the difference in classification. It draws on the knowledge distillation algorithm on the classification task, and treats each pixel as the unit of classification, and performs distillation independently. The goal of pair-wise distillation is to align the pair-wise similarity learned by the simple network (student) and the complex network (teacher), improve the performance of the network through the similarity relationship between two pixels, and use a square error to measure each distillation loss between locations. The overall distillation uses the GAN network, tries to eliminate the difference between the Teacher network and the Student network, and measures the similarity of the overall structure of the image.

3.5 Activation function and loss function

The main function of the activation function is to provide the nonlinear modeling ability of the network. Only after the

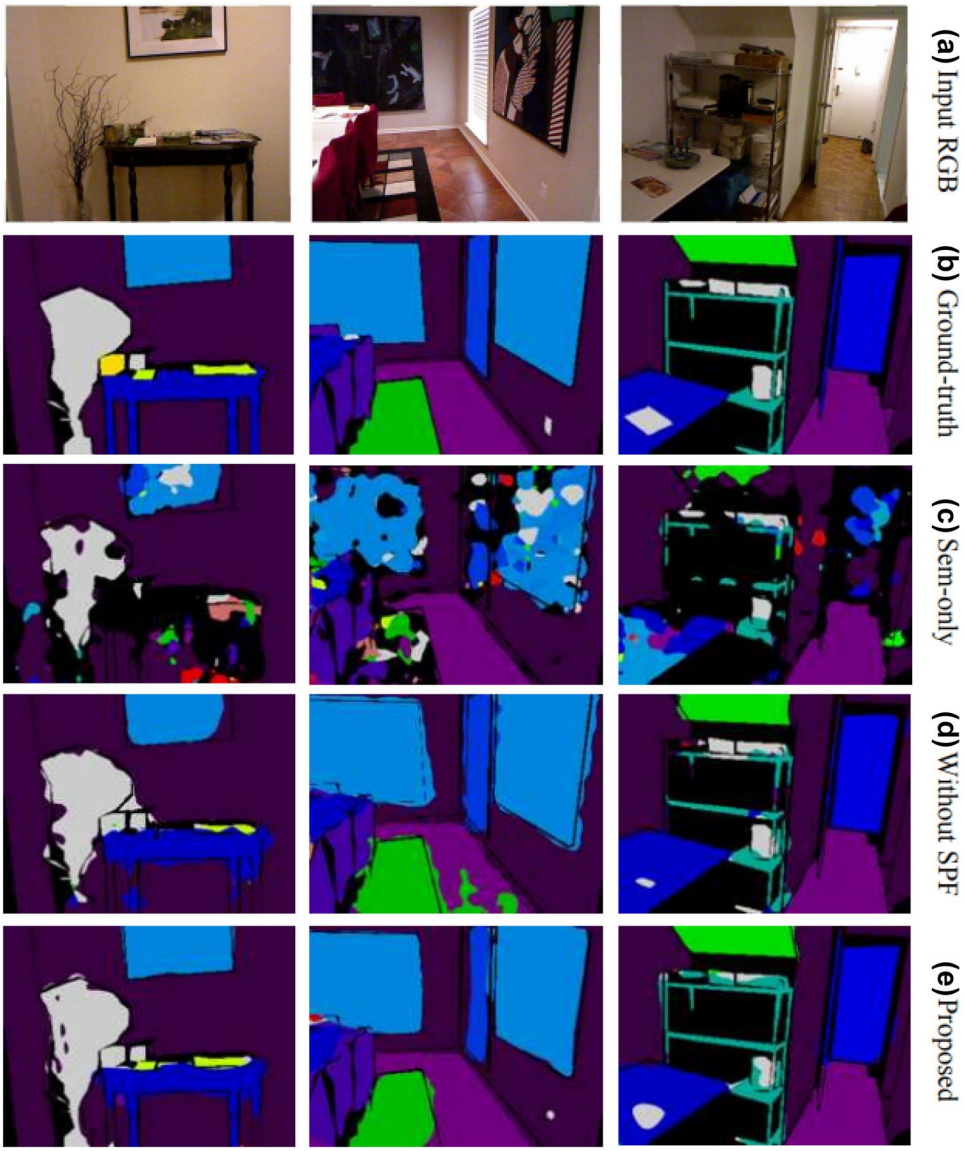

Fig. 12 Jiao et al. proposed method (**e**) is compared to the result with an only semantic branch (**c**) and the result without the SPFs (**d**) [from Jiao et al. (2019)]

activation function is added, could the deep neural network have the hierarchical nonlinear mapping learning ability. The loss function defines the task of the network, solves and evaluates the model by minimizing the loss function. Next, we will introduce the activation function and the loss function in detail.

The activation functions used in semantic segmentation mainly include sigmoid function, tanh function, ReLU function, and its variant functions.

Sigmoid function, also called logistic function, is used for the output of hidden layer neurons, and its value range is (0,1). This function can map a real number to the interval of (0,1), has the shape of a finger function, can be expressed as probability or used for input normalization, typically the sigmoid cross-entropy loss function. The function expression is as follows:

$$S(x) = \frac{1}{1 + e^{-x}} \tag{3}$$

Tanh function is an odd function, and its output mean value is 0, which makes its convergence faster than the sigmoid function and reduces the number of iterations. The function expression is as follows:

$$\tanh x = \frac{\sinh x}{\cosh x} = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} \tag{4}$$

When the characteristics of input data are significantly different, the tanh function is preferred. The sigmoid function is used when the input data features do not differ significantly. Sigmoid and TANH activation function have the same disadvantage: when X is very large or very small,

the gradient is almost zero, so it is very slow to update the network using a gradient descent optimization algorithm.

Relu function is a popular activation function, which is called Rectified Linear Units, and can better alleviate the problem of gradient disappearance. The function expression is:

$$f(x) = \max(0, x) \tag{5}$$

Nowadays, researchers have made some improvements to the activation function in RGBD-based semantic segmentation tasks. For example, Jia et al. 2021) proposed adding spatial regularization to the activation function, which can easily integrate spatial regularization methods such as total variation (TV) into the CNN network. It can help CNN find a better local maximum. The figure of merit makes the segmentation result more robust to noise.

Loss functions include softmax, cross-entropy, focal loss, dice loss.

The cross-entropy loss function checks each pixel separately and compares the class prediction (Softmax or SigmoID) with the object vector (one HOT), including binary cross-entropy and multi-classification cross-entropy. The sigmoid activation function is adopted in the binary cross-entropy function. The last layer of the model contains only one channel. At each pixel, the category probability obtained is p or $1-p$, and the formula is as follows:

$$L = -\left[y \cdot \log(p) + (1-y) \cdot \log(1-p)\right] \tag{6}$$

where y represents the label of the sample, and the positive class is 1 and the negative class is 0. p is the probability that the sample is predicted to be positive.

Multi-classification cross-entropy directly expands the two terms of dichotomies into multinomial terms, and the number of terms is equal to the number of categories. The formula is as follows:

$$L = -\sum_{c=1}^{M} y_c \log\left(p_c\right) \tag{7}$$

where M represents the number of categories; y_c is the indicator variable (0 or 1), 1 if the category is the same as the sample category, 0 otherwise; p_c represents the predicted probability that the observed sample belongs to category c.

Focal Loss is mainly to solve the problem of serious imbalance of positive and negative sample ratio in one-stage target detection. The loss function reduces the weight of a large number of simple negative samples in training and is an improvement based on the cross-entropy loss function. The expression is as follows:

$$FL(p_t) = -(1-p_t)^{\gamma} \log\left(p_t\right)$$

$$p_t = \begin{cases} p & ify = 1 \\ 1-p & otherwise \end{cases} \tag{8}$$

Dice Loss is suitable for extremely uneven samples, and the expression is as follows:

$$d = 1 - \frac{2|X \cap Y|}{|X| + |Y|} \tag{9}$$

$|X|$ and $|Y|$ represent the number of elements in the set X and Y, respectively. In general, the use of Dice Loss will have adverse effects on backpropagation, making the training unstable.

Jiao et al. (2019) extended a loss function for target detection to semantic segmentation tasks to alleviate the problem of data imbalance, while using berHu loss(Eq. 1) for in-depth supervision.

$$L_d = \sum_i \begin{cases} \left|d_i - D_i^*\right|, \left|d_i - D_i^*\right| \le \delta \\ \frac{\left(d_i - D_i^*\right)^2 + \delta^2}{2\delta}, \left|d_i - D_i^*\right| > \delta \end{cases} \tag{10}$$

4 Experiment

In this section, an in-depth investigation is conducted on the evaluation metrics of the RGBD semantic segmentation task and related data sets, and the segmentation accuracy achieved by each method on the two mainstream data sets is also evaluated. The evaluation metrics of semantic segmentation include execution time, memory usage, and accuracy. For different methods and different evaluation objects, the evaluation metrics is also different. For some methods with high real-time requirements, the evaluation metrics are relatively biased towards the execution time and memory usage, to increase the computing speed.

4.1 Data preparation

This part mainly introduces the data sets involved in the RGBD semantic segmentation task. The collection scenarios of the data set are diverse and can be divided into indoor, outdoor, and 3D cloud data sets. Table 1 summarizes the relevant characteristics of some frequently-used data sets, including the data format, training set, validation set, test set, etc.

4.1.1 Indoor data set

The indoor data set mainly includes NYUDv2 (Silberman et al. 2012), SUN-RGBD (Song et al. 2015), Scannet (Dai

Table 1 RGBD semantic segmentation related data set

Dataset	Year	Classes	Resolution	Datatype	Complex/True	Train	Validation	Test
NYUDv2	2012	40	640×480	R,D	R	795	–	654
SUN RGBD	2015	37	Variable	R,D,Pc	R	2666	5050	2619
SID	2017	13	1080×1080	R,D,Pc	R	70,469	–	–
Scannet	2017	21	640×480	R,D	R	19,466	5436	2537
Matterport3D	2017	61	1280×1024	R,D,Pc	R	–	–	–
Semantic3D	2016	8	–	–	R	15个	–	15个
CityScapes	2015	19	2048×1024	R	R	2975	5525	500
KITTI	2012	3/11/10	Variable	R, Pc	R	4000	1000	500
Vaihingen	2012	6	–	Pc	R	–	–	–
SUN3D	2013	33	640×480	Pc	R	19,640	N/A	N/A
MSeg	2020	194	360/720/1080	R	R	190,231		12,561

R RGB, *D* Depth, *Pc* Point cloud

et al. 2017), Stanford Indoor Dataset (SID)(Armeni et al. 2017) and Matterport3D (Chang et al. 2017) data set.

NYUDv2 (Silberman et al. 2012): 2.5-dimensional data set, which contains 1449 indoor RGB-D images captured by Microsoft Kinect equipment. It gives dense pixel-level labels, including category-level and instance-level labels. There are a total of 40 indoor object classes, including 795 in the training set and 654 in the test set. This data set is particularly important because it portrays indoor scenes, making it useful for the task of training certain domestic robots. However, its small scale compared to other data sets limits its application in deep networks.

SUN RGBD (Song et al. 2015): It is captured by four RGB-D sensors, containing 10,000 RGB-D images, the size is the same as PASCAL VOC. The data set contains images in NYU depth v2 (Silberman et al. 2012), Berkeley B3DO (Janoch et al. 2013), and SUN3D (Xiao et al. 2013) data sets. The entire data set is densely annotated, including polygons, bounding boxes with directions, and three-dimensional space, suitable for scenes Understand the task.

SID (Armeni et al. 2017): Contains 70,496 RGB-D images, 13 object categories. It provides a data set of a large indoor space. It also provides various modes of mutual registration from two-dimensional, 2.5d, and three-dimensional domains. It has instance-level semantic and geometric annotations, including registered original and semantically annotated 3D meshes and points clouds. The data set can utilize the existing rules in large indoor spaces to develop joint cross pattern learning models and potentially unsupervised methods.

Scannet (Dai et al. 2017): It is an RGB-D video data set. It contains 19,466 samples for training, 5436 samples for verification, 2537 samples for testing, 1201 scenes for training, and 312 scenes for testing.

Matterport3D (Chang et al. 2017): This is a large-scale indoor RGB-D dataset, which includes 10,800 panoramic views, 194,400 RGB and depth images, and 3D point clouds captured from 90 large-scale scenes. It has two-dimensional and three-dimensional dense annotations of 50,811 object instances, which are mapped to the canonical 40 object classes. The annotations included are surface reconstructions, camera poses, and 2D and 3D semantic segmentation.

4.1.2 Outdoor data set

Outdoor data sets mainly include Semantic3D, CityScapes (Cordts et al. 2016), and KITTI (Uhrig et al. 2017).

Semantic3D: Outdoor scene point cloud database. This database is divided into two types of data: one is semantic8, which contains 8 types of things(man-made terrain, natural terrain, high vegetation, low vegetation, buildings, hardscape, scanning artifacts, cars), there are 15 training sets and 15 test sets each. For the consideration of the algorithm capacity, a compressed version, reduced8, is provided, and the points are much less. The reduce-8 training set is the same as semantic-8, and the test set is only a part of semantic-8. Covers a wide range of urban outdoor scenes: churches, streets, railroad tracks, squares, villages, football fields, and castles.

CityScapes (Cordts et al. 2016): is a large outdoor RGB-D dataset, focusing on the semantic understanding of urban street scenes. Contains pictures of 27 cities, finely annotated 19 semantic class pixel-level labels, composed of about 5000 finely annotated images and 20,000 coarsely annotated images, each with a resolution of 2048 × 1024, of which 2975 images are used For training, 500 images are used for verification and 1525 images are used for testing.

KITTI (Uhrig et al. 2017): It is a widely used outdoor depth estimation data set for autonomous driving and mobile

robots. It includes several hours of traffic scenes recorded with various sensor modes, including high-resolution RGB, grayscale stereo cameras, and 3D laser scanners. There are 4 k images for training, 1 k images for verification, and 500 images for online benchmark testing.

4.1.3 Point cloud data set

The point cloud data set mainly includes the Vaihingen data set and SUN3D (Xiao et al. 2013).

Vaihingen: The data set contains 33 small blocks of different sizes. Each small-block contains a real orthophoto, which is extracted from a larger top mosaic.

SUN3D (Xiao et al. 2013): is a location-centric database, including RGB-D images, camera poses, object segmentation, and point clouds. The dataset contains a large-scale RGB-D video database with 8 annotation sequences. Each frame performs semantic segmentation on the pose information of the objects. It consists of 415 sequences from 41 different buildings in 254 different spaces.

In addition, a composite data set MSeg (Lambert et al. 2020) has been proposed, which combines semantic segmentation data sets in different fields, relabeling more than 220,000 object masks in more than 80,000 images to coordinate classification, and introduce pixel-level annotations into alignment.

4.2 Experimental setting

The execution time of a system or algorithm is very important, which is necessary to meet the real-time performance of the system. In the case of less memory consumption of semantic partition, the same important factor is that the execution time of the same task is better.

The following accuracy metrics are mainly used in the RGBD semantic segmentation task, including pixel accuracy, average pixel accuracy, mean intersection over union, and FWIoU.

Suppose there are k + 1 classes, P_{ij} represents the number of pixels that belong to class i but are predicted to be class j, P_{ii} represents the real number, P_{ij} and P_{ji} are interpreted as false positive and false negative respectively.

Pixel accuracy(PA) is pixel accuracy, which calculates the ratio of the number of correctly classified pixels to the number of all pixels.

$$PA = \frac{\sum_{i=0}^{k} P_{ii}}{\sum_{i=0}^{k} \sum_{j=0}^{k} P_{ij}} \tag{11}$$

Average pixel accuracy(MPA) is mean accuracy. Calculate the ratio of the correct number of pixels in each category to the number of all pixels in that category and then average them. The calculation is based on pixel accuracy.

$$MPA = \frac{1}{k+1} \sum_{i=0}^{k} \frac{P_{ii}}{\sum_{j=0}^{k} P_{ij}} \tag{12}$$

The mean intersection over union (mean-IoU, MIoU) calculates the IoU of each category and then averages the ratio of the intersection and union of the two sets.

$$MIoU = \frac{1}{k+1} \sum_{i=0}^{k} \frac{P_{ii}}{\sum_{j=0}^{k} P_{ij} + \sum_{j=0}^{k} P_{ji} - P_{ii}} \tag{13}$$

Execution time is a valuable indicator in semantic segmentation. However, because the execution time has high requirements on hardware and back-end implementation, researchers are generally required to describe the hardware of the system and the baseline of research methods, so as to facilitate time statistics or comparison. Execution time can help more people evaluate different approaches more equitably. However, in most cases, the importance of execution time is ignored, and people pay more attention to evaluation indexes related to accuracy.

Memory usage is another important indicator in semantic segmentation. In most scenarios, memory can be expandable, but for some embedded devices, memory is precious and means memory is not rich enough. Based on this, when considering the operation effect of the method, the average memory usage can be recorded and the execution conditions can be described more completely, which is of great significance for the subsequent evaluation and comparison of the effect.

Frequency Weighted intersection over union (FWIoU) is an improvement of MIoU. The weight is set according to the frequency of each category, and the IoU of each category is weighted and summed.

4.3 Performance analysis

Most networks choose the NYUDv2 data set and SUN-RGBD data set for training and testing, so in this article, these two data sets are mainly selected to compare various algorithms.

Most methods use the PyTorch framework and SGD optimizer. The crop size is 480 × 640. The comparison results based on the NYUDv2 data set are shown in Table 2 and Fig. 13, and the comparison results based on the SUN RGBD data set are shown in Table 3 and Fig. 14. According to Tables 2 and 3, it can be seen that the current RGBD semantic segmentation effect can reach more than 50% mIoU, and there is still a lot of room for improvement.

Analyzing the different results on the two data sets, it can be seen that the learning method of multimodal fusion can effectively improve the accuracy of semantic segmentation. Most methods choose the ResNet as the backbone, and the

Table 2 Comparison of evaluation results based on NYUDv2 data set

Year	Methods	Backbone	Information	Pixel Acc.(%)	mAcc.(%)	MIoU.(%)
2021	GLPNet	ResNet101	RGB + Depth	79.1	66.6	54.61
	TCDNet	ResNet50	RGB + Depth	77.2	–	52.4
	TCDNet	ResNet101	RGB + Depth	77.8	–	53.1
	Su Y et al	–	RGB + HHA	77.9	–	52.0
	NANet	ResNet50	RGB + Depth	77.1	–	51.4
	NANet	ResNet101	RGB + Depth	77.9	–	52.3
2020	ESANet	ResNet50	RGB + Depth	–	–	50.53
	CANet	ResNet50	RGB + Depth	75.7	62.6	49.6
	CANet	ResNet101	RGB + Depth	76.6	63.8	51.2
	Chen et al	ResNet50	RGB + HHA	77.9	–	52.4
	TSNet	RGB-ResNet34, Depth-VGGNet18	RGB + Depth	73.5	59.6	46.1
	SGNet	ResNet101	RGB + Depth	76.4	62.1	50.3
	MrCNet	–	RGB + Depth	–	56.1	43.1
	MSCFNet	ResNet101	RGB + Depth	–	–	36.0
2019	CTSNet	ResNet101	RGB + HHA	76.3	–	50.6
	ACNet	ResNet50	RGB + Depth	–	–	48.3
	GADNet	ResNet-50	RGB	84.8	68.7	59.6
	GCN	VGG-16	RGB + Depth	69.11	48.49	36.94
	PAPNet	ResNet-50	RGB	76.2	62.5	50.4
	ZZNet	ResNet-152	RGB	77	64.0	51.2
	ZZNet	ResNet-101	RGB	75.8	62.3	49.3
	ICDNet	U-Net	RGB + Depth	–	–	46.05
	KIL	ResNet-101	RGB	75.1	58.4	50.2
	2.5D Conv	ResNet-101	RGB + HHA	75.9	–	49.1
	2.5D Conv	ResNet-101	RGB	75.3	–	48.4
2018	DCNN	VGG-16	RGB + HHA	61.4	35.6	26.2
	DCNN	VGG-16	RGB + Depth	60.3	39.3	27.8
	TRL-jointly	ResNet18	RGB	74.3	55.5	45.0
	TRL-jointly	ResNet50	RGB	76.2	56.3	46.4
2017	RefineNet	ResNet152	–	73.6	58.9	46.5
	RDFNet	2 × ResNet152	RGB + HHA	76	62.8	50.1
	RDFNet	2 × ResNet101	RGB + HHA	75.6	62.2	49.1
	CFNet	RefineNet152	RGB + HHA	–	–	47.7
	3DGNN	VGG16	RGB + HHA	–	55.2	43.1
	LSD-GF	2 × VGG16	RGB + HHA	71.9	60.7	45.9
2015	FCN	2 × VGG16	RGB + HHA	65.4	46.1	34

segmentation effect will be more accurate than that of the VGG network.

Based on the NYUDv2 data set, GADNet (Jiao et al. 2019) achieved the best effect, which reached 59.6% MIoU. It proposed to jointly infer semantic and depth information by extracting geometric perception embedding while mining useful depth domain information, to eliminate this strong constraint. It decoupled the single-task prediction network into two joint tasks of semantic segmentation and geometric embedding learning, which can achieve good results. In 2020, the algorithm proposed by Chen et al. (2020) also reached an MIoU of 52.4%, showing the effectiveness of the attention mechanism and the two-way multi-step communication strategy. Based on the SUN RGBD data set, the best effect is also GADNet (Jiao et al. 2019). In addition, the proposal of KIL (2019) can also make the semantic segmentation effect reach 52.0% mIoU, showing the effectiveness of multi-task joint learning.

For the network that only uses RGB as the only input, the final segmentation effect is sometimes better than the effect of additional depth input (Jiao et al. 2019; Zhang et al. 2019; Lin and Huang 2019; Zhou et al. 2019), and the effectiveness of the multi-task joint learning(Zhang et al. 2019; Zhou et al. 2019) can be seen. For the network whose input

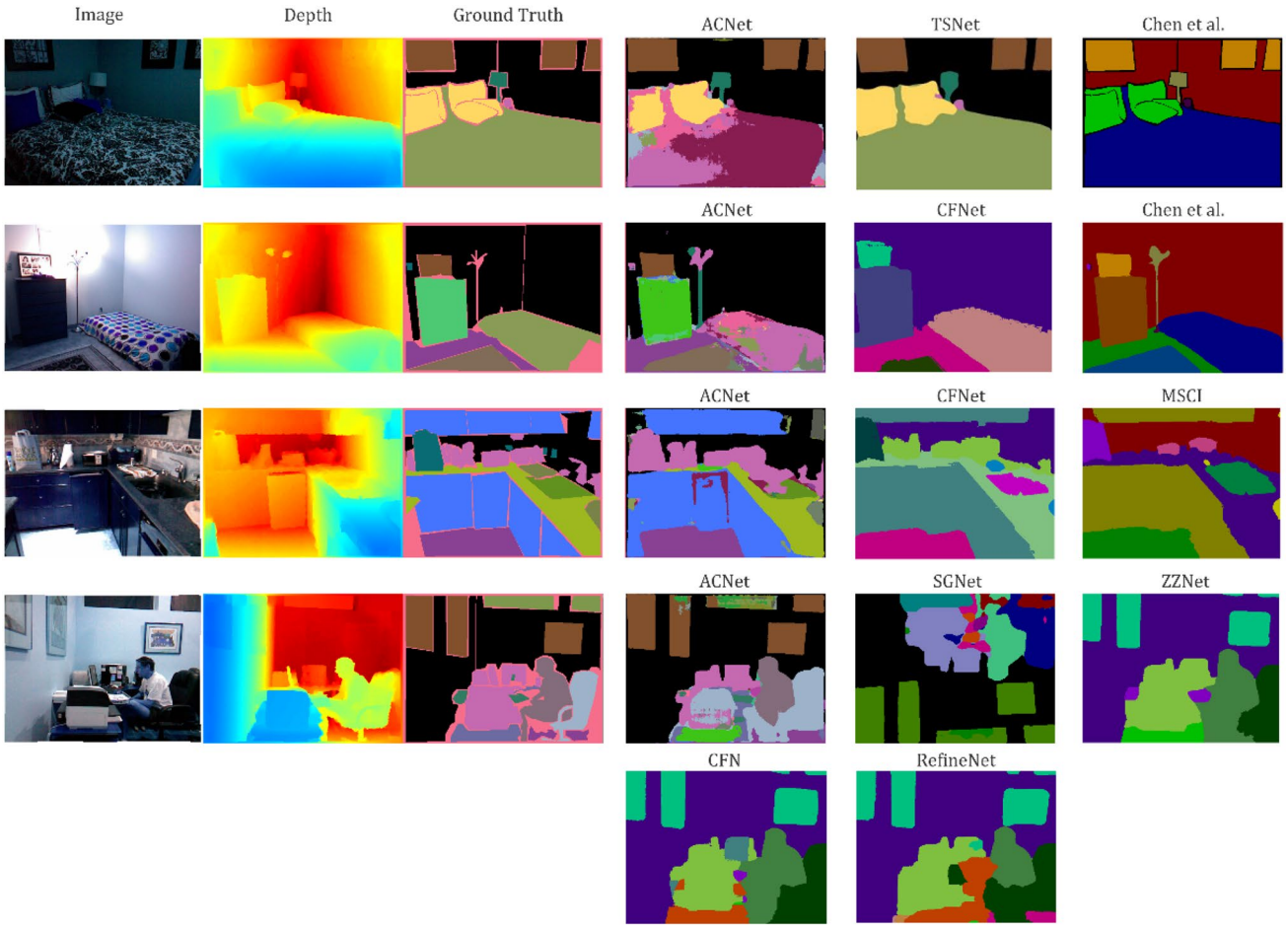

Fig. 13 Sample of the comparison to state-of-the-art models. Scene images are taken from the NYUDv2 dataset (Silberman et al. 2012)

is RGB and depth data, HHA can achieve a better segmentation effect, and the network whose input is DEPTH and RGB has lower segmentation accuracy. That is to say, HHA contains more abundant information, which means that we will have richer and more detailed annotations for DEPTH in the data set in the future. HHA is horizontal disparity, height above ground, and the angle the pixel's local surface normal makes with the inferred gravity direction. However, the HHA coding method only emphasizes the complementary information between the data of each channel and ignores the independent components of each channel, which has certain limitations.

At the same time, most algorithms only disclose the results of pixel accuracy and mIoU. Data about execution time and memory usage are not disclosed. On the one hand, this phenomenon is because most algorithms only focus on the improvement of accuracy, but ignore the measurement of execution time and memory occupation, and ignore the requirements of these two indicators in real life. On the other hand, there is no uniform indicator for measuring execution time and memory usage, and the baseline settings of each experiment are also different, which can't be completely unified. The papers and related codes involved are summarized in the links https://github.com/krqh/RGBD-semantic-segmentation-review.git.

Table 4 introduces the RGBD semantic segmentation method in the FPS of backbone. The larger the FPS value is, the better the real-time performance is. As can be seen from the table, the real-time performance of RGBD semantic segmentation network is also improving.

5 Discussion and conclusion

RGBD semantic segmentation has been applied in many fields such as autonomous driving and robotics, and the related algorithms are becoming more and more abundant. The current algorithm trend is an effective fusion of traditional methods and deep learning methods. For the application of RGBD semantic segmentation, the algorithm will pay more attention to real-time performance, and the source

Table 3 Comparison of evaluation results based on SUNRGBD data set

Year	Methods	Backbone	Information	Pixel Acc.(%)	mIoU.(%)	mAcc.(%)
2021	GLPNet	ResNet101	RGB + Depth	82.8	51.2	63.3
	TCDNet	ResNet50	RGB + Depth	82.9	48.8	–
	TCDNet	ResNet101	RGB + Depth	83.1	49.5	–
	Su Y et al	–	RGB + HHA	81.8	50.6	–
	NANet	NANet	RGB + Depth	82.1	48.0	–
	NANet	NANet	RGB + Depth	82.3	48.8	–
2020	ESANet	ResNet50	RGB + Depth	–	48.31	–
	CANet	ResNet50	RGB + Depth	81.6	48.1	59.0
	CANet	ResNet101	RGB + Depth	82.5	49.3	60.5
	Chen et al	ResNet50	RGB + HHA	82.5	49.4	–
	SGNet	ResNet101	RGB + Depth	81.8	48.5	60.9
	MrCNet		RGB + Depth	–	42.8	54.6
	MSCFNet	VGG-16	–	–	26.94	–
	MSCFNet	ResNet101	–	–	33.56	44.2
2019	ACNet	ResNet50	RGB + Depth	–	48.1	–
	CTSNet	ResNet101	RGB + HHA	82.4	48.5	–
	GADNet	ResNet-50	RGB	85.5	54.5	74.9
	PAPNet	ResNet-50	RGB	83.8	50.5	58.4
	ZZNet	ResNet-152	RGB	84.7	51.8	62.9
	ZZNet	ResNet-101	RGB	82.7	48.6	61.3
	KIL	ResNet-50	RGB	84.0	51.5	57.6
	KIL	ResNet-101	RGB	84.8	52.0	58.0
	2.5D conv	ResNet-101	RGB + HHA	82.4	48.2	–
	2.5D conv	ResNet-101	RGB	81.9	47.3	–
2018	RedNet	ResNet-34	RGB + Depth	80.8	46.8	58.3
	RedNet	ResNet-50	RGB + Depth	81.3	47.8	60.3
	DCNN	VGG-16	RGB + Depth	72.4	29.7	38.6
	DCNN	VGG-16	RGB + HHA	72.9	31.3	41.2
	TRL-jointly	ResNet18	RGB	81.1	46.3	56.3
	TRL-jointly	ResNet50	RGB	83.6	49.6	58.2
2017	RDFNet	ResNet-152	RGB + HHA	76.0	50.1	62.8
	CFNet	RefineNet152	RGB + HHA	–	48.1	–
2016	FuseNet	VGG-16	RGB + HHA	–	34.02	50.07

of the depth image will more and more conform to the nonlinear law. Based on the above research, the future research directions of RGBD semantic segmentation are listed below:

(1) More effective multi-task learning: decoupling a single task into a combination of multiple tasks, the depth information required in RGBD semantic segmentation can be obtained from the depth estimation task, combined with RGB images. The local and global fusion of depth information and RGB information can be more effective, which makes the segmentation result more robust. Find a better loss function to adapt to multi-task learning, and find a more suitable method of fusing high-level and low-level features to better improve the RGBD semantic segmentation results;

(2) Data augmentation: The existing data set is still not rich enough. In the future, you can try to combine static images and video to form a larger data set and do a good job of classification and calibration. It is believed that in the future, the collection of in-depth information will achieve better results. At the same time, it can also combine more point cloud data sets to better perform RGBD semantic segmentation;

(3) Real-time: In future work, the focus can be on improving the calculation speed of the algorithm, simplifying the complex operation, to better meet the practical needs and perform better and faster segmentation. The common ideas include reducing the redundant convolution core in the network, increasing the pre-data processing, reducing the complex fusion mode, reducing the number of channels, using the appropriate loss function, and so on;

(4) Memory consumption: There are already some networks that can be lightweight so that they can be suitable for more devices. Knowledge distillation is an effective

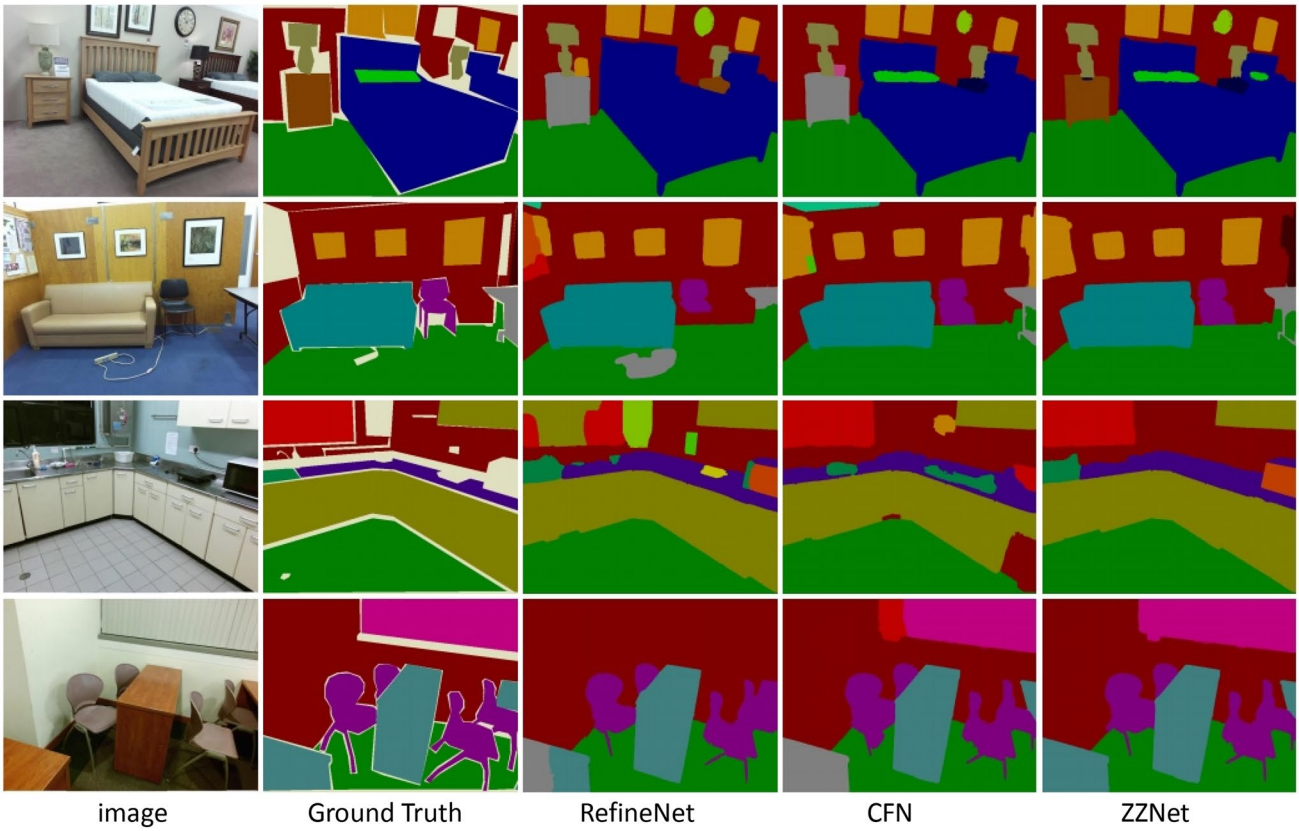

Fig. 14 Sample of the comparison to state-of-the-art models. Scene images are taken from the SUN-RGBD dataset (Song et al. 2015) [from ZZNet (Lin and Huang 2019)]

Table 4 Time performance display of segmented network

Network	Backbone	FPS
FCN	2×VGG16	8
RefineNet	ResNet152	16
RDFNET	2×ResNet152	9
RDFNET	2×ResNet101	11
3DGNN	VGG16	5
D-CNN	2×VGG16	13
D-CNN	VGG16	26
ACNet	2×ResNet50	18
ACNet	3×ResNet50	16.5
SGNet	ResNet101	28
RedNet	2×ResNet34	26
RedNet	2×ResNet50	22.1
SA-GATE	2×ResNet50	11.9
ESANet	2×ResNet34	27.5
ESANet	2×ResNet50	22.6

method. At the same time, the number of parameters is reduced by other means to achieve the purpose of reducing memory consumption. It can be expanded from the aspects of simplification of algorithms and lightweight software.

This article mainly summarizes the relevant algorithms of RGBD semantic segmentation, describes the feature fusion, development history, and key technologies, and summarizes and analyzes the existing RGBD semantic segmentation methods, which provides the foundation and extension direction for future research work. Firstly, various fusion methods of RGB and depth features are introduced, and the existing RGBD semantic segmentation methods are classified according to their development history. At the same time, the key points that can effectively improve the segmentation accuracy are described in detail, and the data set used is sorted out. In addition, we compare the accuracy of these methods on NYUD and SUN RGBD data sets, and analyze the advantages and disadvantages of these methods, which have certain significance for future work.

References

Armeni I, Sax S. Zamir AR, Savarese S (2017) Joint 2d-3d-semantic data for indoor scene understanding. https://doi.org/10.48550/arXiv.1702.01105

Badrinarayanan V, Kendall A, Cipolla R (2017) Segnet: a deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans Pattern Anal Mach Intell 39(12):2481–2495

Chang A, Dai A, Funkhouser T, Halber M, Niessner M, Savva M, Zhang Y (2017) Matterport3d: learning from rgb-d data in indoor environments. arXiv preprint arXiv:1709.06158

Chen LZ, Lin Z, Wang Z, Yang YL, Cheng MM (2021a) Spatial information guided convolution for real-time RGBD semantic segmentation. IEEE Trans Image Process 30:2313–2324

Chen X, Lin K Y, Wang J, Wu W, Qian C, Li H, Zeng G (2020, August) Bi-directional cross-modality feature propagation with separation-and-aggregation gate for RGB-D semantic segmentation. In: European conference on computer vision. Springer, Cham, pp 561–577

Chen S, Zhu X, Liu W, He X, Liu J (2021b) Global-local propagation network for RGB-D semantic segmentation. arXiv preprint arXiv:2101.10801

Cheng Y, Cai R, Li Z, Zhao X, Huang, K (2017) Locality-sensitive deconvolution networks with gated fusion for rgb-d indoor semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3029–3037

Cordts M, Omran M, Ramos S, Rehfeld T, Enzweiler M, Benenson R, 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

Couprie C, Farabet C, Najman L, LeCun, Y (2013) Indoor semantic segmentation using depth information. arXiv preprint arXiv:1301.3572

Dai A, Chang AX, 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

Deng L, Yang M, Li T, He Y, Wang C (2019) RFBNet: deep multimodal networks with residual fusion blocks for RGB-D semantic segmentation. arXiv preprint arXiv:1907.00135

Gao X, Yu J, Li J (2019, July) RGBD semantic segmentation based on global convolutional network. In: Proceedings of the 2019 4th international conference on robotics, control and automation, pp 192–197

Giannone G, Chidlovskii B (2019) Learning common representation from RGB and depth images. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, pp 0–0

Gupta S, Arbeláez P, Girshick R, Malik J (2015) Indoor scene understanding with rgb-d images: bottom-up segmentation, object detection and semantic segmentation. Int J Comput Vision 112(2):133–149

Gupta S, Girshick R, Arbeláez P, Malik, J (2014) Learning rich features from RGB-D images for object detection and segmentation. In: European conference on computer vision Springer, Cham, pp 345–360

Hazirbas C, Ma L, Domokos C, Cremers D (2016) Fusenet: incorporating depth into semantic segmentation via fusion-based cnn architecture. In: Asian conference on computer vision. Springer, Cham, pp 213–228

He K, Zhang X, Ren S, Sun J (2015) Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Trans Pattern Anal Mach Intell 37(9):1904–1916

He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778

He Y, Chiu WC, Keuper M, Fritz M (2017) Std2p: Rgbd semantic segmentation using spatio-temporal data-driven pooling. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4837–4846

Hinton G, Vinyals O, Dean J (2015) Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531, 2(7)

Hu X, Yang K, Fei L, Wang K (2019) Acnet: attention based network to exploit complementary features for rgbd semantic segmentation. In: 2019 IEEE international conference on image processing (ICIP). IEEE, pp 1440–1444

Janoch A, Karayev S, Jia Y, Barron JT, Fritz M, Saenko K, Darrell T (2013) A category-level 3d object dataset: putting the kinect to work. In: Consumer depth cameras for computer vision. Springer, London, pp 141–165

Jia F, Liu J, Tai XC (2021) A regularized convolutional neural network for semantic image segmentation. Anal Appl 19(01):147–165

Jiang J, Zhang Z, Huang Y, Zheng L (2017) Incorporating depth into both cnn and crf for indoor semantic segmentation. In: 2017 8th IEEE international conference on software engineering and service science (ICSESS). IEEE, pp 525–530

Jiang J, Zheng L, Luo F, Zhang Z (2018) Rednet: residual encoder-decoder network for indoor rgb-d semantic segmentation. arXiv preprint arXiv:1806.01054

Jiao J, Wei Y, Jie Z, Shi H, Lau RW, Huang TS (2019) Geometry-aware distillation for indoor semantic segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 2869–2878

Kosiorek A (2017) 神经网络中的注意力机制. 机器人产业, 6

Lambert J, Liu Z, Sener O, Hays J, Koltun V (2020) MSeg: a composite dataset for multi-domain semantic segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, p 2879–2888

Li Z, Gan Y, Liang X, Yu Y, Cheng H, Lin L (2016) Lstm-cf: unifying context modeling and fusion with lstms for rgb-d scene labeling. In: European conference on computer vision. Springer, Cham, p 541–557

Li Y, Zhang J, Cheng Y, Huang K, Tan T (2017) Semantics-guided multi-level RGB-D feature fusion for indoor semantic segmentation. In: 2017 IEEE international conference on image processing (ICIP), pp 1262–1266. IEEE.

Lin D, Huang H (2019) Zig-zag network for semantic segmentation of RGB-D images. IEEE Trans Pattern Anal Mach Intell 42(10):2642–2655

Lin X, Sánchez-Escobedo D, Casas JR, Pardàs M (2019) Depth estimation and semantic segmentation from a single RGB image using a hybrid convolutional neural network. Sensors 19(8):1795

Lin D, Chen G, Cohen-Or D, Heng PA, Huang H (2017a) Cascaded feature network for semantic segmentation of RGB-D images. In: Proceedings of the IEEE international conference on computer vision, pp 1311–1319

Lin G, Milan A, Shen C, Reid I (2017b) Refinenet: multi-path refinement networks for high-resolution semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1925–1934

Lin D, Ji Y, Lischinski D, Cohen-Or D, Huang H (2018) Multi-scale context intertwining for semantic segmentation. In: Proceedings of the European conference on computer vision (ECCV), pp 603–619

Liu H, Wu W, Wang X, Qian Y (2018a) RGB-D joint modelling with scene geometric information for indoor semantic segmentation. Multimed Tools Appl 77(17):22475–22488

Liu J, Wang Y, Li Y, Fu J, Li J, Lu H (2018b) Collaborative deconvolutional neural networks for joint depth estimation and semantic segmentation. IEEE Trans Neural Netw Learning Syst 29(11):5655–5666

Liu Y, Chen K, Liu C, Qin Z, Luo Z, Wang J (2019) Structured knowledge distillation for semantic segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 2604–2613

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

McCormac J, Handa A, Leutenegger S, Davison AJ (2016) Scenenet rgb-d: 5m photorealistic images of synthetic indoor trajectories with ground truth. arXiv preprint arXiv:1612.05079.

Nakajima Y, Kang B, Saito H, Kitani K (2019) Incremental class discovery for semantic segmentation with RGBD sensing. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 972–981

Park SJ, Hong KS, Lee S (2017) Rdfnet: Rgb-d multi-level residual feature fusion for indoor semantic segmentation. In: Proceedings of the IEEE international conference on computer vision, pp 4980–4989

Qi X, Liao R, Jia J, Fidler S, Urtasun R (2017) 3d graph neural networks for rgbd semantic segmentation. In: Proceedings of the IEEE international conference on computer vision, pp 5199–5208

Schneider L, Jasch M, Fröhlich B, Weber T, Franke U, Pollefeys M, Rätsch M (2017) Multimodal neural networks: Rgb-d for semantic segmentation and object detection. In: Scandinavian conference on image analysis Springer, Cham, pp 98–109

Seichter D, Köhler M, Lewandowski B, Wengefeld T, Gross HM (2021) Efficient rgb-d semantic segmentation for indoor scene analysis. In: 2021 IEEE international conference on robotics and automation (ICRA). IEEE, pp 13525–13531

Shi W, Zhu D, Zhang G, Chen L, Wang L, Li J, Zhang X (2019) Multilevel Cross-Aware RGBD Semantic Segmentation of Indoor Environments. In: 2019 IEEE international conference on cyborg and bionic systems (CBS). IEEE, pp 346–351

Silberman N, Hoiem D, Kohli P, Fergus R (2012) Indoor segmentation and support inference from rgbd images. In: European conference on computer vision, Springer, Berlin, Heidelberg, pp 746–760

Song S, Lichtenberg SP, Xiao J (2015) Sun rgb-d: a rgb-d scene understanding benchmark suite. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 567–576

Su W, Wang Z (2016) Regularized fully convolutional networks for RGB-D semantic segmentation. In: 2016 visual communications and image processing (VCIP). IEEE, pp. 1–4

Su Y, Yuan Y, Jiang Z (2021) Deep feature selection-and-fusion for RGB-D semantic segmentation. In: 2021 IEEE international conference on multimedia and expo (ICME) IEEE, pp 1–6

Sun L, Yang K, Hu X, Hu W, Wang K (2020) Real-time fusion network for RGB-D semantic segmentation incorporating unexpected obstacle detection for road-driving images. IEEE Robot Autom Lett 5(4):5558–5565

Uhrig J, Schneider N, Schneider L, Franke U, Brox T, Geiger A (2017, October) Sparsity invariant cnns. In: 2017 international conference on 3D Vision (3DV) IEEE, pp 11–20

Wang Y, Chen Q, Chen S, Wu J (2020b) Multi-scale convolutional features network for semantic segmentation in indoor scenes. IEEE Access 8:89575–89583

Wang W, Neumann U (2018) Depth-aware cnn for rgb-d segmentation. In: Proceedings of the European conference on computer vision (ECCV), pp 135–150

Wang J, Wang Z, Tao D, See S, Wang G (2016) Learning common and specific features for RGB-D semantic segmentation with deconvolutional networks. In: European conference on computer vision. Springer, Cham, pp 664–679

Wang G, Wang Z, Chen Y, Wang G, Chen J (2020) Indoor scene semantic segmentation based on RGB-D image and convolution neural network. J Phys Conf Ser 1637(1):012138

Xiao J, Owens A, Torralba A (2013) Sun3d: a database of big spaces reconstructed using sfm and object labels. In: Proceedings of the IEEE international conference on computer vision, pp 1625–1632

Xing Y, Wang J, Chen X, Zeng G (2019a) 2.5 D convolution for RGB-D semantic segmentation. In: 2019a IEEE international conference on image processing (ICIP). IEEE, pp 1410–1414

Xing Y, Wang J, Chen X, Zeng G (2019b) Coupling two-stream RGB-D semantic segmentation network by idempotent mappings. In: 2019b IEEE international conference on image processing (ICIP). IEEE, pp 1850–1854

Yue Y, Zhou W, Lei J, Yu L (2021) Two-stage cascaded decoder for semantic segmentation of RGB-D images. IEEE Signal Process Lett 28:1115–1119

Zhang G, Xue JH, Xie P, Yang S, Wang G (2021) Non-local aggregation for RGB-D semantic segmentation. IEEE Signal Process Lett 28:658–662

Zhang Z, Cui Z, Xu C, Jie Z, Li X, Yang J (2018) Joint task-recursive learning for semantic segmentation and depth estimation. In: Proceedings of the European conference on computer vision (ECCV), pp 235–251

Zhang Z, Cui Z, Xu C, Yan Y, Sebe N, Yang J (2019) Pattern-affinitive propagation across depth, surface normal and semantic segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 4106–4115

Zhen M, Wang J, Zhou L, Fang T, Quan L (2019) Learning fully dense neural networks for image semantic segmentation. Proc AAAI Conf Artif Intell 33(1):9283–9290

Zheng Z, Xie D, Chen C, Zhu Z (2020) Multi-resolution cascaded network with depth-similar residual module for real-time semantic segmentation on RGB-D images. In: 2020 IEEE international conference on networking, sensing and control (ICNSC). IEEE, pp 1–6

Zhou L, Xu C, Cui Z, Yang J (2019) KIL: knowledge interactiveness learning for joint depth estimation and semantic segmentation. In: Asian conference on pattern recognition, Springer, Cham, pp 835–848

Zhou H, Qi L, Wan Z, Huang H, Yang X (2020a) RGB-D Co-attention network for semantic segmentation. In: Proceedings of the Asian conference on computer vision

Zhou W, Yuan J, Lei J, Luo T (2020b) TSNet: Three-stream self-attention network for RGB-D indoor semantic segmentation. IEEE Intell Syst 36(4):73–78