www.nature.com/scientificreports

scientific reports

OPEN

Multi-scale fusion for RGB-D indoor semantic segmentation

Shiyi Jiang[1], Yang Xu[1,2]✉, Danyang Li[1] & Runze Fan[1]

In computer vision, convolution and pooling operations tend to lose high-frequency information, and the contour details will also disappear with the deepening of the network, especially in image semantic segmentation. For RGB-D image semantic segmentation, all the effective information of RGB and depth image can not be used effectively, while the form of wavelet transform can retain the low and high frequency information of the original image perfectly. In order to solve the information losing problems, we proposed an RGB-D indoor semantic segmentation network based on multi-scale fusion: designed a wavelet transform fusion module to retain contour details, a nonsubsampled contourlet transform to replace the pooling operation, and a multiple pyramid module to aggregate multi-scale information and context global information. The proposed method can retain the characteristics of multi-scale information with the help of wavelet transform, and make full use of the complementarity of high and low frequency information. As the depth of the convolutional neural network increases without losing the multi-frequency characteristics, the segmentation accuracy of image edge contour details is also improved. We evaluated our proposed efficient method on commonly used indoor datasets NYUv2 and SUNRGB-D, and the results showed that we achieved state-of-the-art performance and real-time inference.

Along with the development of the computer vision field, in view of the image semantic segmentation has become an important subject in the field of image segmentation is to classify each pixel in the image and predict belongs to tags, location, through the image is divided into several of the same areas of the nature of the process to provide a complete understanding of the scenario[1]. At present, semantic segmentation has been widely applied in automatic driving[2], remote sensing analysis[3], medical image processing[4], etc. As for the semantic segmentation of indoor scenes, the factors affecting the semantic segmentation of indoor scenes are quite complex (such as illumination, occlusion, etc.)[5]. Since common downsampling operations of neural networks such as Max pooling, average pooling, and convolution operations do not separate noise. In the study of indoor scenes, the main purpose is to reduce the influence of these factors and improve the accuracy of semantic segmentation. Previous deep learning approach by dealing with the indoor scene RGB image to achieve end-to-end image semantic segmentation, due to the complex indoor scene, uneven illumination, color texture repeat degrees higher, the indoor scene semantic based on RGB image segmentation is category error classification, edge false segmentation, robustness and accuracy is not high. In recent years, research has found that semantic segmentation based on RGB-D images can improve the segmentation effect through the depth information in the scene which is less affected by lighting and other conditions. At the same time, the depth information can also reflect the position and other relationships between objects, which is complementary to RGB color images and also has a certain auxiliary role in semantic segmentation. Couprie et al.[6] found that adding depth information could improve the segmentation accuracy of similar objects.

With the emergence of depth cameras such as TOF (Time-of-flight) and Kinect, it becomes easier to obtain the depth information of scenes. However, it is always a challenging problem to find an effective and high-quality fusion method between RGB color images and depth information. Jiang et al.[7] added and fused the RGB and depth information of the middle layer of the encoder through parallel processing. Li et al.[8] believed that single-layer fusion could not complement the color image and depth map well, and chose to fuse features before final prediction. Chang et al.[9] added depth information to the loss function and used the change of depth value on the edge of the classified object to constrain the network training. However, there are still many problems with the above approach: Factors effecting on the one hand, the indoor noise exists in different frequency part of image more[10], while traditional convolution neural network down sampling operations such as average pooling, maximum pooling will not separate different frequency information, which can lead to high frequency noise increased with the increase of the depth of the network are preserved, the sampling data of aliasing in the basic structure of the residual noise destroys the image features, Thus, it brings difficulties to the image segmentation

[1]College of Big Data and Information Engineering, Guizhou University, Guiyang 550025, China. [2]Guiyang Aluminum-magnesium Design and Research Institute Co., Ltd., Guiyang 550009, China. ✉email: xuy@gzu.edu.cn

task. On the other hand, the depth image is used as the fourth channel to fuse with the color image, which does not make full use of the complementarity of RGB color information and depth information. With the increase of depth of neural network, there are problems such as multi-scale information feature loss in different downsampling methods. However, for the problem of indoor scene semantic segmentation, the efficiency of multi-scale feature extraction affects the segmentation accuracy of small target objects. Compared with the RGB image, the depth image is more sparse, which represents the depth value of each pixel in the RGB image. The edge contour information, that is, high-frequency information, can be better obtained from the depth image, but traditional convolution, pooling and other operations often lose high-frequency features. The direct splicing fusion method requires convolution operation to add a large number of calculation parameters, and the network model is too large for the neural network to become an urgent problem.

In the process of signal processing, in order to avoid aliasing features and separate information of different frequencies, time-frequency analysis tools are usually used to decompose data into different frequency intervals, such as wavelet transform[11]. In recent years, combining wavelet transformation with deep learning approach has developed rapidly, such as whether Liu et al.[12] wavelet transform and inverse transformation to replace the network of sampling for super-resolution images, or Ramamonjisoa et al.[13], such as using wavelet transform fusion into the network in the middle tier RGB color image processing, good results have been obtained. But simply replace wavelet transform in the network under the sampling operation did not make full use of it in the face of non-stationary information can be effectively extracted features characteristic of different scales, different frequency domain information is also different, for visual reasoning tasks, different from the traditional sampling retention under the low frequency information, high frequency information is often contains the detail[14] the outline of the object, especially for indoor RGB-D image segmentation, noise, depth and other effects exist in different scale and frequency information. This is a big problem that cannot be solved by the above methods. Aiming at the problems existing in the RGB-D semantic segmentation method of indoor scenes. Inspired by the above methods, this paper proposes a semantic segmentation network for indoor scene RGB-D images based on multi-scale fusion (MSFNet), aiming at the existing problems in indoor scene RGB-D semantic segmentation methods, The network encoder uses Resnet-50 as the baseline, and designs a fusion Module based on wavelet transform. The original image is divided into four frequency children by wavelet transform, and image fusion is realized by inverse wavelet transform, which is respectively used for RGB color image and depth image fusion. In the deep layer of the network, the feature information of different scales is fused, and the information of different scales of the image is more refined. And different from ASPP and other related fusion modules, the wavelet transform fusion module does not add additional calculation, which is more lightweight. And the refinement on the feature resolution is more intensive and accurate. We evaluated our network on commonly used indoor datasets NYUv2[15] and SUNRGB-D[16] and obtained high-quality results. The main contributions of this paper are as follows:

- We propose a semantic segmentation network for indoor scene RGB-D images based on multi-scale fusion (MSFNet), aiming at the existing problems in indoor scene RGB-D semantic segmentation methods.
- A new multi-scale fusion method is proposed. Through wavelet transform and inverse wavelet transform, images of subbands with different frequencies are fused, Nonsubsampled contour wave transform replaces the pooling operation of the encoder in baseline, preserving the multi-scale properties and directionality of the original image, and designed a multiple pyramid module (MPM) to aggregate multi-scale information and context global information, retained the edge contour information of images, the number of parameters is reduced, and the network operation efficiency is accelerated.
- Extensive experiments using NYUv2 and SUN-RGBD public datasets demonstrate that our proposed method has better performance and robustness than the current mainstream methods in different data domains.

Related work

Semantic segmentation. In the field of computer vision, the emergence of deep learning has made up for many deficiencies in traditional methods. In 2015, Long et al.[17] proposed Fully Convolutional Networks (FCN), which replaced the last Fully connected layer of the traditional convolutional neural network with a deconvolution layer, realized the "end-to-end" RGB color image semantic segmentation, and added the pooling layer and jump connection. Address feature loss issues. Ronneberger et al.[18] proposed U-NET, which adopted an Encoder-decoder structure for small targets and added richer feature information fusion. Badrinarayanan et al.[19] proposed Segmentation Network (SegNet), which adopted the same structure as U-NET. In order to prevent information loss, SegNet adopted pooling with index to solve the problem of location information loss caused by multiple pooling. Based on ResNet[20], Zhao et al.[21] proposed the Pyramid Scene Parsing Network (PSPNet), which introduces the Pyramid fusion module to integrate the feature information of different scales according to the prior knowledge of the context in the Scene, and solves the problem of spatial information loss in FCN. There are also recent approaches to improve segmentation accuracy by adding supervision during training, Borse et al.[22] propose Hierarchically Supervised Semantic Segmentation (HS3), a training scheme that supervises intermediate layers in a segmentation network to learn meaningful representations by varying task complexity. Common network architectures for semantic segmentation follow an encoder-decoder design: the encoder extracts features from the input and downsamples them to reduce computational effort, the decoder upsamples, deconvolves them to recover the input resolution, and finally assigns a semantic class to each input pixel. However, there are also recent studies that combine CNN with Transformer. Li et al.[23] proposed a dual encoding-decoding structure of the X-shaped network, integrated both characteristics of CNN and Transformer, achieves good segmentation results

RGB-D semantic segmentation. The depth image can be used as the complement of RGB color image to provide the geometric information of the scene, so as to improve the accuracy of segmentation. However, how effectively integrate deep information into network training is still a challenge. The fusion method can be classified into three types: early fusion, middle fusion and late fusion. Some of the early methods directly the depth image to the RGB color channels[7,24,25], and under the assumption that there are four channel input RGB-D data in the training, but the depth of information as a color image directly between the fourth channel is not very good complementarity, training and the use of two branches of the network, one for RGB color images, One is used for depth information image, and the two are fused in the middle layer to get good results. In this way, each branch can extract its own features and use them for fusion, such as color and texture from RGB images, geometry, and location information independent of lighting from depth images. Hazirbas et al.[26] proposed Fuse-Net, which used two branches to extract features from RGB and depth images at the same time, and fused the depth features into RGB feature maps with the deepening of the network. Hu et al.[27] proposed ACNet to fuse the features extracted from RGB and depth images on the third encoder branch and added the attention module. In the middle of fusion, people begin to pay attention to the fusion in different stages, and explore different fusion methods. Gupta et al.[28] proposed the Horizontal dimensions, Height above ground, Angle of the surface normal (HHA) depth information representation. The depth image is converted into different channels (horizontal difference, ground height and surface normal vector Angle), but HHA only emphasizes the complementary information between the data of each channel but ignores its independence, and it is computatively expensive. At present, more and more research focuses on changing the fusion efficiency and using different levels of information. Xing et al.[29] propose a novel method to effectively integrate RGB and HHA features By replacing identity mappings in Resnet-based two-stream network with idempotent Mappings. Chen et al.[30] proposed Spatial Information Guided Convolution (S-CONV), which effectively integrates RGB features with 3D Spatial Information. However, in the process of feature fusion, both convolution operation and pooling operation are accompanied by the loss of feature information, and the complementarity of RGB color image and depth image cannot be fully utilized. Most existing methods exploit a multi-stage fusion strategy to propagate depth feature to the RGB branch. However, at the very deep stage, the propagation in a simple element-wise addition manner can not fully utilize the depth information. Chen et al.[31] proposed Global-Local propagation network (GLPNet) to solve this problem. they used a local context fusion module (L-CFM) to dynamically align both modalities before element-wise fusion, and designed a global context fusion module (G-CFM) to propagate the depth information to the RGB branch by jointly modeling the multi-modal global context features. Cao et al.[32] introduced a Shape-aware Convolutional layer (ShapeConv) for processing the depth feature, where the depth feature is firstly decomposed into a shape-component and a base-component, next two learnable weights are introduced to cooperate with them independently, and finally a convolution is applied on the re-weighted combination of these two components.. This operation need to consume large amounts of computing resources, however, is not conducive to deploy on mobile equipment, combined with the RGB-D image fusion problem, according to the characteristics of the wavelet transform, we will be combined with neural network, wavelet transform for image processing and fusion of the different frequencies by multi-scale method to improve the effect of color image and the depth of the image fusion, And because of the characteristics of wavelet transform, it does not consume additional computing resources.

Wavelet transform in computer vision. Wavelet transform is often used in signal processing and image analysis because of its multi-resolution analysis and stepwise decomposition. At the same time, the multi-scale decomposition of wavelet transform is more consistent with human vision mechanism. In neural networks, both convolution operation and pooling operation (maximum pooling, average pooling) are lost in different frequency information to a certain extent, and the information features of different frequencies can be retained by combining with wavelet transform. Bae et al.[33] combined wavelet transform with residual network and found that more subbands of wavelet transform could improve the learning effect of network. Guo et al.[34] proposed deep Wavelet Super Resolution to improve network performance by processing the missing details in the process of subband restoration. Li et al.[35] proposed to replace the pooling operation in neural network with wavelet transform, which could retain the high-frequency information and edge details of the original image. Li et al.[36] used wavelet transform and inverse wavelet transform instead of downsampling and upsampling operations in U-Net. Ramamonjisoa et al.[37] integrate wavelet transforms into the encoder-decoder process, requiring less than half the multiply adds in the decoder network. Wavelet transform instead of down sampling operation, however, does not take full advantage of image multi-scale, frequency characteristics, more is not necessary to use wavelet transform to simply reduce the resolution of the image, using the wavelet transform can keep the different scales of information fusion, especially in the depth of the image and high frequency and low frequency characteristics of the color image is not the same, Based on the characteristics of the above wavelet transform, we propose a multi-scale fusion RGB-D indoor image segmentation network, which integrates the low-frequency and high-frequency features of the original image without adding extra computation.

Proposed method

In this paper, we adopted the “encoder-decoder” structure, and designed the wavelet transform fusion module. The nonsubsampled contourlet wave transform is used to replace the pooling operation and the inverse wavelet transform to improve the semantic segmentation accuracy. The encoder network adopts Resnet-50 as the backbone network, and removes the fully connected layer. There are two branches in the network to extract the RGB and depth feature information in the original image respectively. At the same time, the feature fusion of RGB-D is carried out through the wavelet transform fusion module. At the connection between the encoder and the decoder, we designed a multiple pyramid module to aggregate feature information and global context information

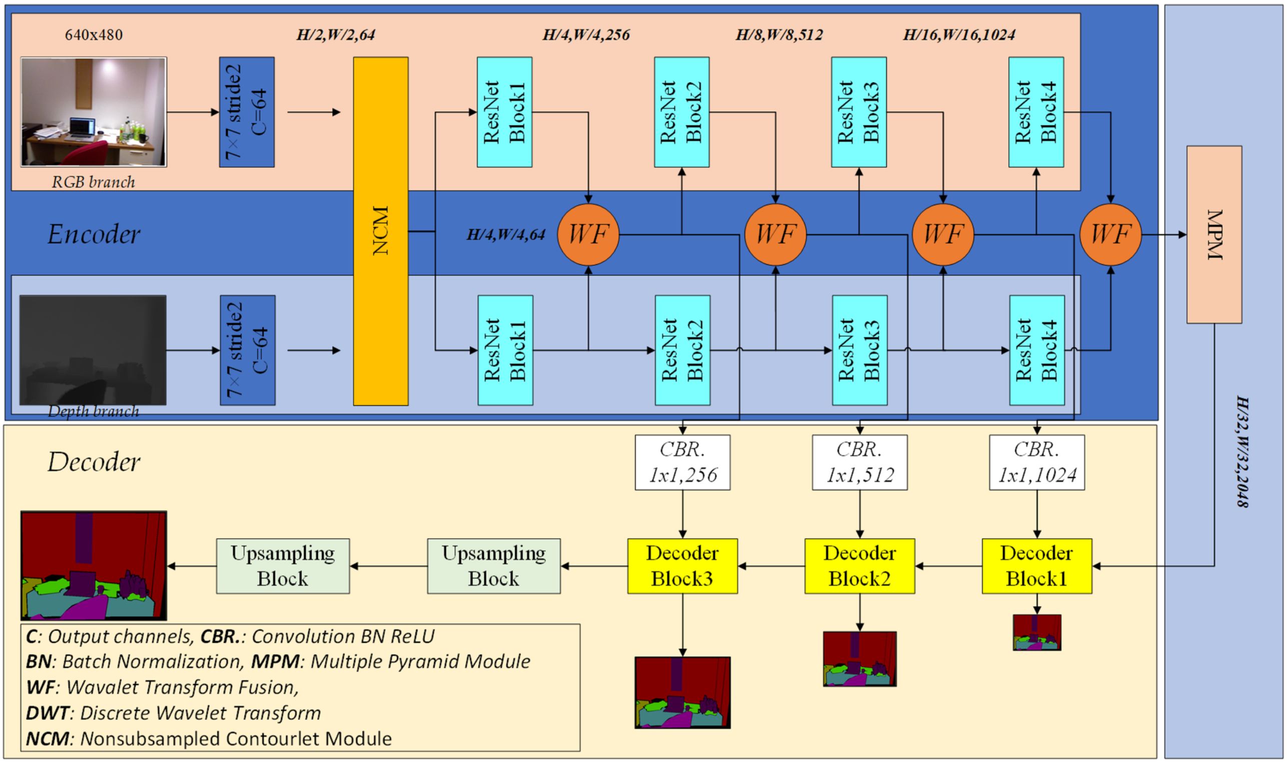

Figure 1. Network structure of model.

at different scales. Finally, the above features are up-sampled several times through the decoder network. Each module of the decoder upsamples the features by a factor of two, and performs better feature mapping by convolution and connection of the encoder. Then, the high-resolution images are gradually recovered and the semantic segmentation results are output. The network model structure is shown in Fig. 1.

Wavelet transform.

In this part, we first introduce 2D Wavelet Transform. We select the efficient and fast Haar Wavelet, then introduce the designed Wavelet Transform fusion module.

The original image X can be decomposed into four subband images by using the discrete wavelet transform (DWT)[38] of 2D Haar wavelet transform, it can decompose the original image into four subband images, and the image size, that is, the image resolution, becomes half of the original image. The above operation can be equal to the decomposition of the original image x using four filters (f_{LL}, f_{LH}, f_{HL}, f_{HH}),obtained four subband image x_{LL}, x_{LH}, x_{HL}, x_{HH}, namely low frequency A, vertical detail image V, horizontal detail image H and diagonal detail image D. The parameters of the filter are fixed, and they are not updated by the gradient descent operation along with the model training. The filter of Haar wavelet is shown in Eq. (1).

$$f_{LL}=\begin{bmatrix}1&1\\1&1\end{bmatrix},f_{LH}=\begin{bmatrix}-1&-1\\1&1\end{bmatrix},f_{HL}=\begin{bmatrix}-1&1\\-1&1\end{bmatrix},f_{HH}=\begin{bmatrix}1&-1\\-1&1\end{bmatrix} \tag{1}$$

The input image is $x(i,j)$, i represents row and j represents column. The 2D DWT is shown below:

$$\begin{cases} A = x_{LL} = f_{LL} \bigotimes x = x(2i-1,2j-1) + x(2i-1,2j) \\ \quad +x(2i,2j-1) + x(2i,2j) \\ V = x_{LH} = f_{LH} \bigotimes x = -x(2i-1,2j-1) - x(2i-1,2j) \\ \quad +x(2i,2j-1) + x(2i,2j) \\ H = x_{HL} = f_{HL} \bigotimes x = -x(2i-1,2j-1) + x(2i-1,2j) \\ \quad -x(2i,2j-1) + x(2i,2j) \\ D = x_{HH} = f_{HH} \bigotimes x = x(2i-1,2j-1) - x(2i-1,2j) \\ \quad -x(2i,2j-1) + x(2i,2j) \end{cases} \tag{2}$$

In Eq. (2), $\bigotimes$ represents a convolution operation, the input x can be represented by a convolution operation with a different filter, and it can also be understood as downsampling with a stride of 2. Since the wavelet transform does not lose information, the wavelet transform and the inverse wavelet transform are reversible operations. For the Haar wavelet, the inverse operation can be expressed as Eq. (3).

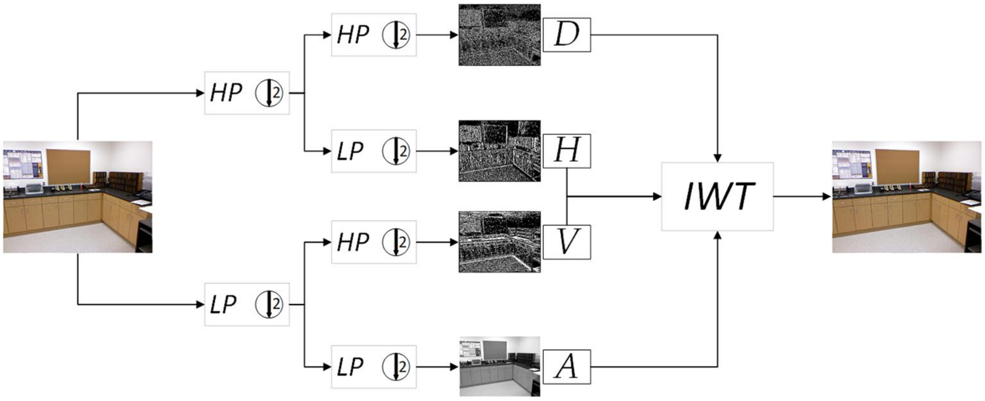

Figure 2. Discrete wavelet transform (DWT) and inverse wavelet inverse transform (IWT).

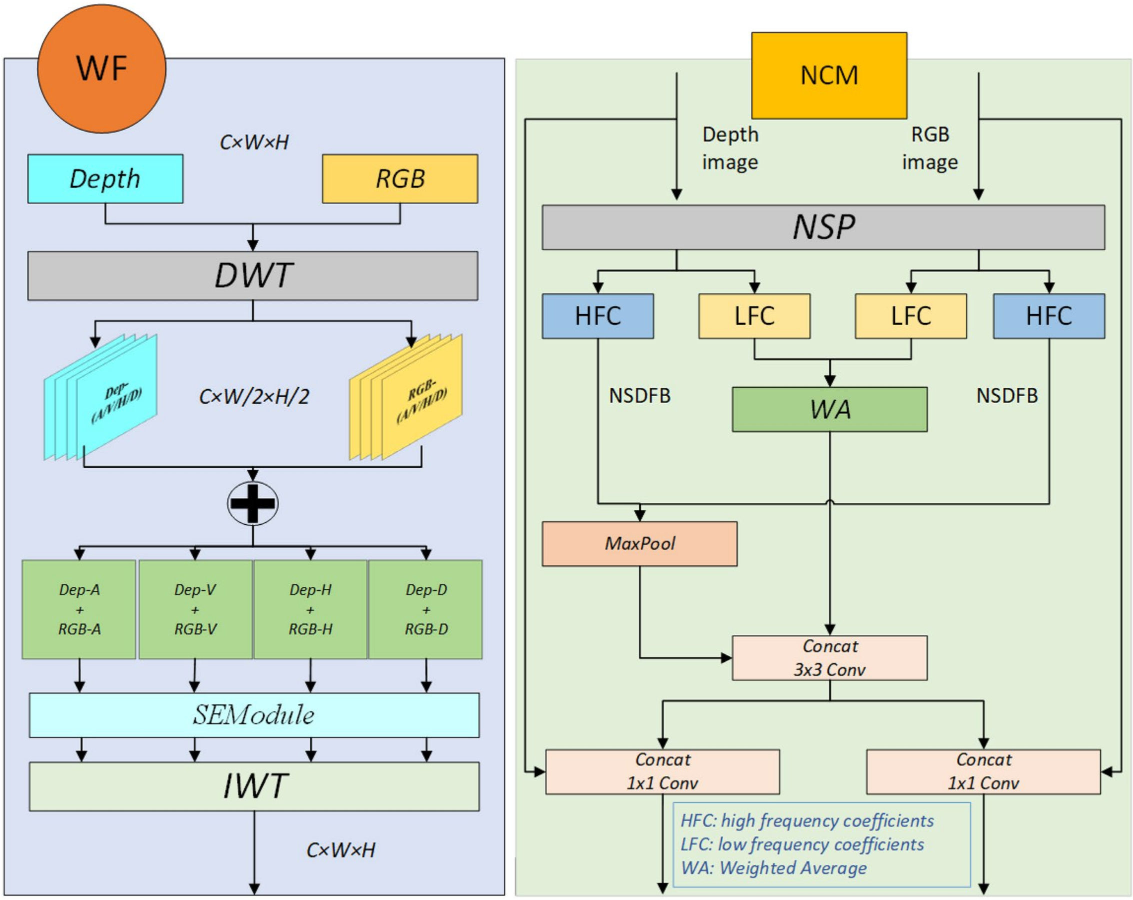

Figure 3. Relevant module.

$$
\begin{cases} x(2i-1,2j-1) = (x_{LL} - x_{LH} - x_{HL} + x_{HH})/4 \\ x(2i-1,2j) = (x_{LL} - x_{LH} + x_{HL} - x_{HH})/4 \\ x(2i,2j-1) = (x_{LL} + x_{LH} - x_{HL} - x_{HH})/4 \\ x(2i,2j) = (x_{LL} + x_{LH} + x_{HL} + x_{HH})/4 \end{cases} \tag{3}
$$

The original image obtains four sub-bands through DWT, or the four sub-bands obtain the original image through IWT. The process diagram is shown in Fig. 2. HP represents high-pass filtering, and LP represents lowpass filtering. ↓2 means the standard downsampling operator with factor 2.

The modules covered with Wavelet fusion and nonsubsampled contourlet module are shown in Fig. 3.

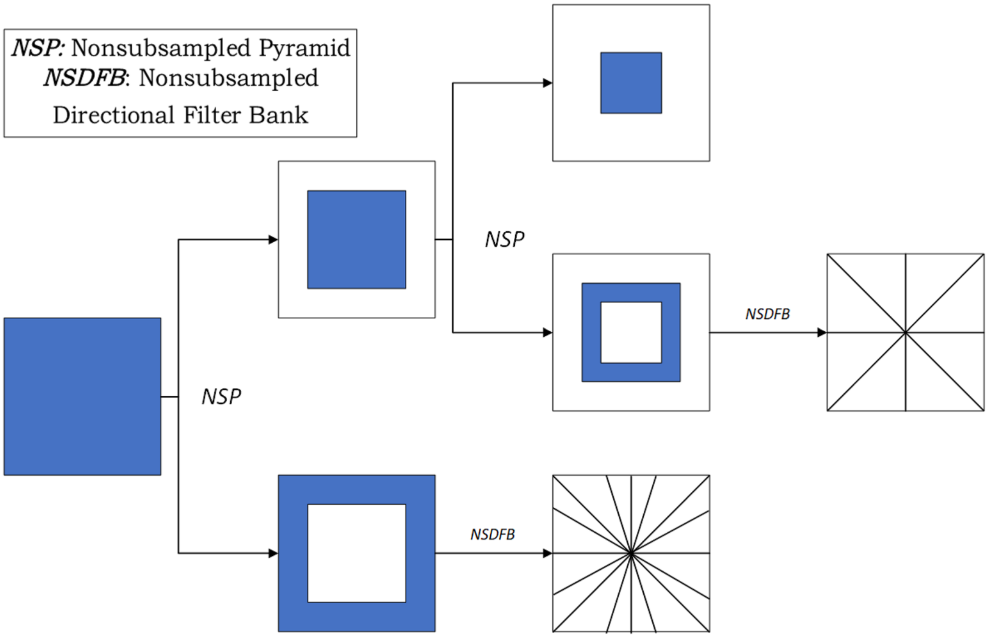
NSP: Nonsubsampled Pyramid
NSDFB: Nonsubsampled Directional Filter Bank
NSP
NSP
NSDFB
NSDFB

Figure 4. Structure of nonsubsampled contourlet transform.

Wavelet fusion module. The Wavelet transform fusion module is added to the encoder to fuse RGB color information and depth information. The two feature information are first decomposed into four different subbands (A, V, H, D) by DWT, and then the corresponding subbands of RGB color image and depth image are added. Through the attention module (SEModule)[39], the output is finally fused by the inverse wavelet transform.

DWT block. When the image information enters the block, it is decomposed into four subbands through DWT, and then it concats the four subbands. The resolution of the obtained feature information becomes half of the input, and the number of channels become 4 times, as a result the number of channels becomes the same as the input channel value through a 1 × 1 convolution operation.

Nonsubsampled contourlet module. The nonsubsampled contourlet module (NCM) is designed to replace the pooling operation in ResNet. We find that the pooling operation in the neural network will lose the corresponding information in the downsampling process more or less, and the features learned by a single convolution operation are limited. At the same time, another disadvantage of pooling operation is that the neural network will lose translation invariance. To solve the above problems, we designed NCM, combined with convolution operation, The Nonsubsampled Pyramid (NSP) and Nonsubsampled Directional Filter Bank (NSDFB) preserve the multi-scale properties and directionality of the original image. This is particularly important for the fusion of depth information. Nonsubsampled Contourlet transform as shown in Fig. 4.

The decomposition scale of the image is defined as D. NSP can obtain the subband information with the number of $D + 1$ by decomposition. The decomposition formula is as follows:

$$R = \sum_{i=0}^{i-1} 2^{k_l} \tag{4}$$

The number of decomposition levels is denoted by i, and the number of decompositions is an integer power of two.The Nonsubsampled contourlet module is shown in the following equation:

$$I = i_j + \sum_{i=0}^{i-1} \sum_{k=0}^{l_i} d_{j,k} \tag{5}$$

The low-frequency self-band information decomposed in the j scale direction is denoted by i_j, and the subband information in the k direction is denoted by $d_{j,k}$

Multiple pyramid module. Atrous Spatial Pyramid Pooling (ASPP) can expand the receptive field, strengthen the contact different context information, but as a result of dilated convolution is a layer of the neighboring pixels from independent subset is obtained by convolution operation, lack of dependence on each other, and the convolution results from a layer of the independent set of local information is missing, so we designed the Multiple pyramid module, as shown in Fig. 5. Firstly, channel splitting is carried out for the current input feature map $F_{in} \in \mathbb{R}^{HxWxC}$, which is divided into N parts. Then, each part $F_{in} \in \mathbb{R}^{HxWx\frac{C}{n}}$ is subjected to Atrous Spatial Pyramid Pooling (ASPP), and the obtained feature map is splicing. Thus, the situation of ignoring the internal information relation while processing a single feature map is avoided, and the global context information feature is fully extracted. ASPP includes a 1x1 convolution, three 3x3 convolution which dilated rate respec-

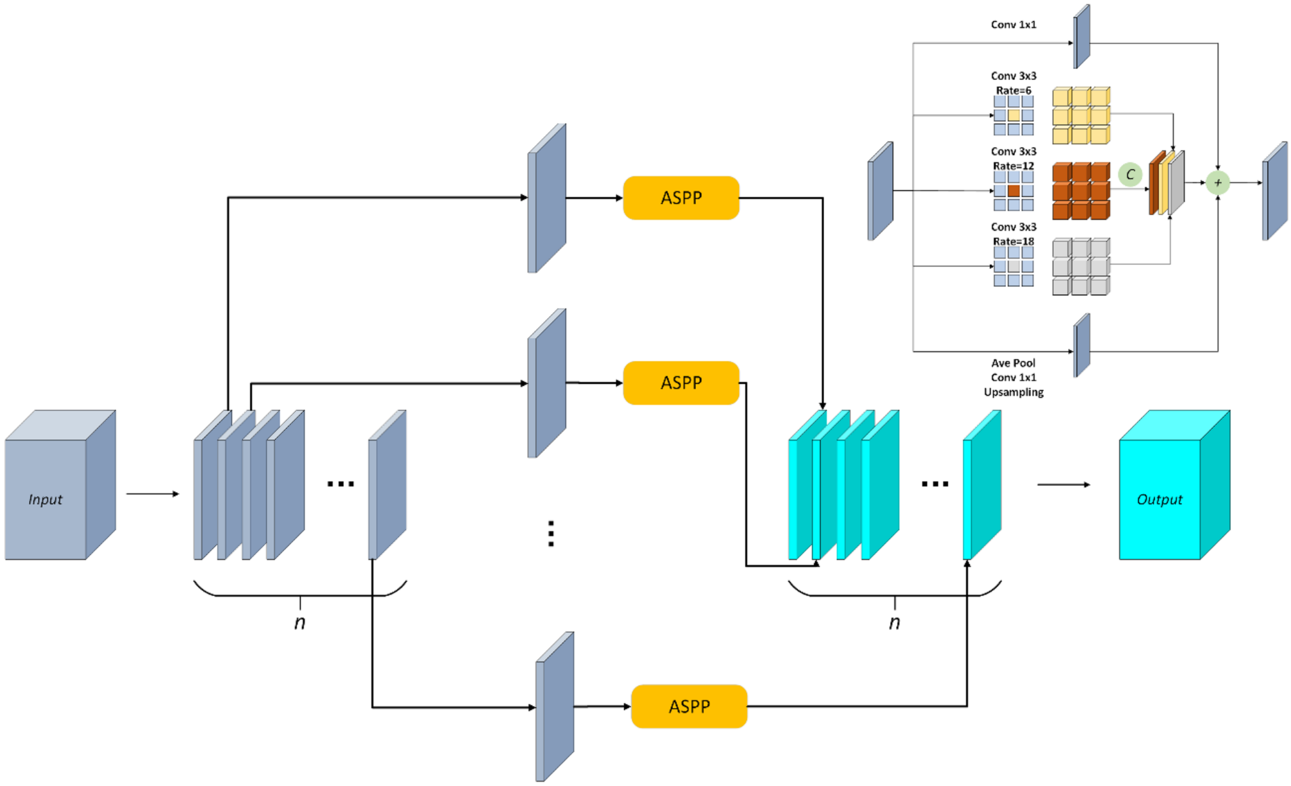

Figure 5. Structure of multiple pyramid module.

tively for 6, 12, 18 and the average pooling, we use concatenation operation on empty part of the convolution operation, keep more original location information, using the way of combined operation 1x1 convolution and average characteristics after the pooling, which ensure every dimension contains more information. The dilated convolution operation is as follows:

$$y(i,j) = \sum_{u=0}^{H} \sum_{v=0}^{W} x(i + ar \times u, j + ar \times v) \times Weight(u,v) \quad (6)$$

where H,W represents the length and width of the input image, *x(i,j)* represents the pixel value at the position *(i,j)*, AR represents the cavity rate, y represents the output, *(u,v)* represents the central coordinate of the convolution kernel, and *Weight* represents the global average pooling of the weight of the convolution kernel at the corresponding position, as shown below:

$$X'_{(i,j)} = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} X_{(i,j)} \quad (7)$$

$X_{(i,j)}$ represents the pixel value of the input image at position *(i,j)*,$X'_{(i,j)}$ represents the pixel value of the output feature map.

Encoder. Resnet-50 is adopted as the backbone network of the encoder, and there are four ResNet blocks in each of the two branches. At the same time, for high-frequency information such as prominent edge contour represented in the depth map, as the depth of the encoder in the network gradually deepens, some edge contour features may be lost. Therefore, we add a discrete wavelet transform module to replace the pooling operation in the original ResNet to retain more information of different frequencies, and fuse the high-frequency edge information obtained from the depth image into the RGB color image through the wavelet transform fusion module.

Decoder. The decoder in this paper is composed of three decoder modules. The decoder module is shown in Fig. 6., and the number of channels is reduced from 2048 with the resolution increasing through convolution operation. In addition, we integrated deep separable convolution[40]. Deep separable convolution can reduce the number of parameters and save the computational cost. Through these operations, the accuracy and efficiency of semantic segmentation can be further improved. Finally, the resolution is increased by nearest neighbor upsampling, and the feature information is integrated by depth-separable convolution. It is difficult to avoid information loss during the up-sampling process of the decoder. We use 1 × 1 convolution operation to integrate the fused features of the encoder RGB-D through decoder hopping connection, and restore the image resolution through two up-sampling modules after three decoder modules. At the same time, we designed a multi-level Loss function, adding an output after each decoder module, inputting outputs of different resolutions and final

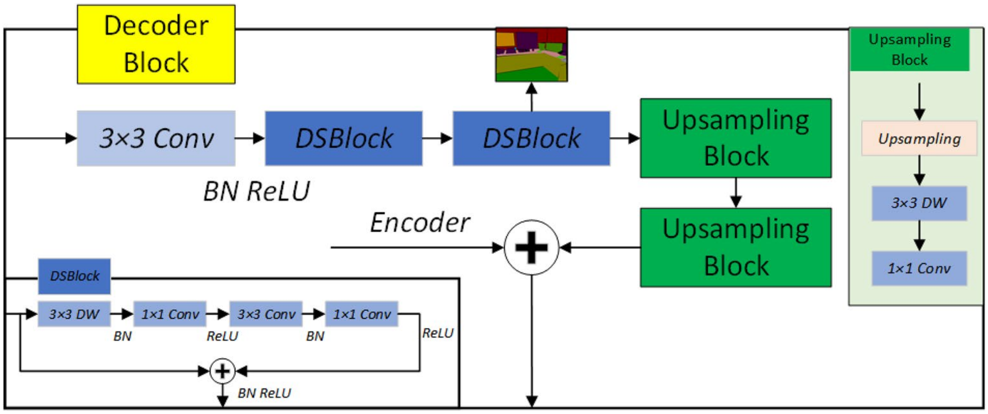

Figure 6. Decoder module.

results to the end of the network and finally obtaining the Loss function. The Loss function selects the cross-entropy function, as shown in Eq. (6):

$$\text{Loss}(x, \text{class}) = \frac{1}{N}\sum_i -\log\left(\frac{\exp(x[\text{class}])}{\sum_j \exp(x[j])}\right) \tag{8}$$

where class represents the category of pixel *i* in the label graph, *X* represents the score of pixel *i*, and *N* represents the total resolution of the output image. Since there are also three decoder module outputs, the total loss function is the sum of four partial loss functions, and because different outputs have different resolutions, we assign different weights according to the size of the resolution, and the ratio is 1:2:3:4. Algorithm 1 shows the details of the proposed method.

```
Algorithm 1: Multi-scale fusion for RGB-D Indoor Semantic Segmentation
Input: Data: original rgb images X, depth images Y, source domain labels $C$, suppose x $\in$ X, y $\in$ Y, c $\in$ $C$.
Output: Predicted labels of the target: $C_{x,y}$
The network architecture consists of encoder, decoder and MPM module in the middle;
while iteration is effective do
    x and y pass through the NCM module to obtain the output $O_1$ according to equation(5);
    foreach x and y pass through ResNet Block do
        x and y are fused by wavelet transform through WF module
    end
    The rgb image $x$ and depth image $y$ are passed through the network encoder to obtain the output $O_2$;
    foreach $MPM(O_2)$ do
        $X_{(i,j)} \in O_2$;
        $X'_{(i,j)} = \frac{1}{H \times W}\sum_{i=1}^{H}\sum_{j=1}^{W} X_{(i,j)}$ $O_3 \leftarrow MPM(O_2)$;
        $X_{(i,j)}$represents the pixel value of the input image at position (i,j),$X'_{(i,j)}$represents the pixel value of the output
          feature map.;
    end
    ($O_4$ , four loss parts) $\leftarrow$ $Decoder(O_2)$;
    $Loss \leftarrow$ Compare$(O_4, c)$ by equation(8)
end
```

Experiments and results

In this section, we first introduce the parameters and environment settings of the experiment, and then introduce the dataset and the selected evaluation metrics. Then, ablation experiments are designed for each module, such as the fusion module and the multiple pyramid module, to verify the effect of the designed module and the rationality of the hyperparameters. Finally, the corresponding experimental results are displayed on two public datasets.

Environment of experiment. In order to verify the performance of the proposed method, this paper conducts experiments on two public datasets, NYUV2 and SUN-RGBD datasets. RGB color images and depth images are input at the same time. The environment of this experiment is with Intel I7-12700 CPU and 64G memory and an NVIDIA GeForce RTX 3090 graphics card. The model is trained by gradient descent method,

Method	Attention	MIoU (%)	FPS
Baseline + RedNet	None	48.30	31
Baseline + ACNet	ACM	49.28	17
Baseline + ESANet	SE	51.05	21
MSFNet	SE	52.23	24.7

Table 1. Ablation experiment of fusion module.

and the deep learning framework of PyTorch is selected at the same time. The learning rate is set to $5e-4$, and the optimizer selects Range optimizer to update the parameters. At the same time, transfer learning[41] is used to take Resnet-50 pre-training weight on Imagenet dataset as the initial weight, which effectively accelerates the training speed of the entire network. The epoch of training is set to 300, and save the weight with best result of validation set.

Datasets and performance measures. We design experiments on two public datasets, NYUV2 and SUN RGB-D, the NYUv2 dataset contains 1449 densely annotated RGB image and depth image pairs, including a total of 464 different scenes. In addition, the dataset contains 35064 different objects, covering 894 different object categories in 1449 images. Our work adopts the standard division strategy of NYUv2 dataset: 795 images are used as the train set, and the remaining 654 images are used as the test set. We also used the common 40-class label setting SUNRGB-D dataset has 37 categories, including 10335 RGBD images with densely labeled semantic labels. The dataset consists of several existing small-scale datasets and some RGB-D images taken by the authors themselves, including 3784 images (taken with Kinect v2), 1159 images (taken with Intel RealSense), 1449 images of NYU v2 dataset (taken with Kinect v1), 554 images carefully selected from Berkeley B3DO dataset (taken with Kinect v1), and 3389 images selected from SUN3D dataset (taken with Asus Xtion). It uses 5050 images as the testset and 5285 images as the trainset. The ablation experiment and semantic segmentation effect test in our work are based on NYUv2, and the entire network input resolution is 480 × 640.

Three semantic segmentation evaluation indexes were used to evaluate the experimental results in this paper, including Pixel Accuracy (PA), Mean Pixel Accuracy, MPA and Mean Intersection over Union, and MIoU. PA is the ratio of the number of correctly classified pixels and all pixels in a picture, equation is defined as follows:

$$PA = \frac{\sum_{i=0}^{k} p_{ii}}{\sum_{i=0}^{k} \sum_{i=0}^{k} p_{ij}} \tag{9}$$

where p_{ii} represents the number of pixels that are correctly classified, and p_{ij} represents the number of pixels that belong to class i but are predicted to be class j.

MPA is the average value of the ratio between the number of correctly classified pixels in each class and all pixel points, which is defined as follows:

$$MPA = \frac{1}{k} \times \sum_{i=0}^{k} \frac{p_{ii}}{\sum_{j=0}^{k} p_{ij}} \tag{10}$$

k indicates the number of categories.

MIoU is the average value of the ratio of the intersection and union of two sets of real value and predicted value, which is defined as follows Eq. (9):

$$MIoU = \frac{1}{k} \times \sum_{i=0}^{k} \frac{p_{ii}}{\sum_{j=0}^{k} p_{ij} + \sum_{j=0}^{k} p_{ji} - p_{ii}} \tag{11}$$

Ablation experiments. In this subsection, we design a series of experiments to evaluate the effectiveness of each module, and the ablation experiments are performed on the NYUv2 dataset.

Ablation experiment of fusion module. In this section, we design multiple experiments to compare the wavelet transform fusion module with other fusion methods. RedNet adds depth image and RGB image directly. ACNet extracts image features by balancing the distribution of features through ACM (Attention Complementary Modules) and adding a third branch. The SE (Squeeze and Excitation) module is used first for feature extraction, followed by the additive fusion. Except for the fusion mode, the other modules remain unchanged, and the experimental results are shown in the Table 1.According to the experimental results, it can be seen that the Wavelet transform fusion has a good performance in segmentation accuracy and inference speed, thanks to the characteristics of the wavelet transform without additional computation and the effective retention of the contour details of the original image. In the fusion part, different from other methods, the original image is decomposed into four sub-bands through wavelet transform and inverse wavelet transform, so that each sub-band can retain its own frequency characteristics, and the feature performance is strengthened through the attention mechanism, and then the performance in the depth map and color map is highlighted.

Ablation experiment of nonsubsampled contourlet transform module. In order to verify the effectiveness of the Nonsubsampled contourlet transform module we designed, we designed A set of comparison experiments.

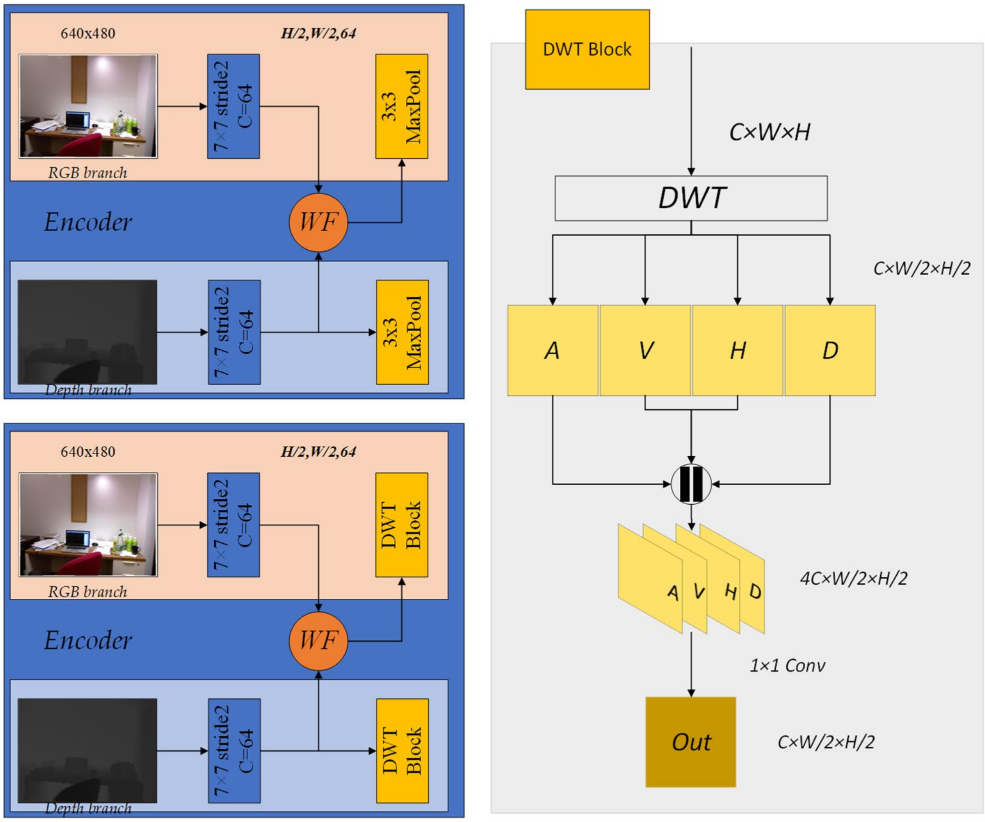

Figure 7. Compare with pooling operations.

Method	Pooling operation	MIoU (%)
Model A	3 × 3 Max pool	51.66
Model B	DWT block	51.85
MSFNet	NCM	52.23

Table 2. Ablation experiment of nonsubsampled contourlet transform module.

Model A used ResNet original pooling operation, namely 3x3 maximum pooling, and used wavelet transform to fuse the feature maps of the two branches. Model B uses wavelet transform to concatenate the four subbands to obtain the down-sampling result. The operation process is shown in Fig. 7, and the result is shown in Table 2.

It can be seen from the results that the segmentation accuracy has been slightly improved after adding NCM. Although the wavelet transform module has also been improved, there are still some deficiencies in the extraction and retention of edge contour features in the shallow layer of the network due to the lack of translation sensitivity and multi-direction recognition by simply using wavelet transform. NCM retains the multi-scale characteristics and directionality of the original image more effectively, Preserving the above features in the thousand-layer part of the neural network is particularly important for the whole training process.

Ablation experiment of multiple pyramid module. The hyperparameter experiments on the multiple pyramid module are tested on the NYUv2 dataset. In this section, we study the influence of hyperparameter n on the performance of the fusion module. N means that the feature map is divided into several parts according to the number of channels. Five groups of experiments with N of 2, 4, 6, 8 and 10 are designed, and the evaluation indexes are mIoU and inference speed, the result is shown in Fig. 8.

As can be seen from the above result, with the gradual increase of N, mIoU also increases, reaching 52.23% when N is 6. However, the intersection ratio fluctuates with the subsequent increase of N, and the inference speed gradually decreases. On this basis, we select N = 6 as the hyperparameter of the multi-pyramid module according to the inference speed and segmentation accuracy.

In order to verify the effectiveness of the multi-pyramid module we designed, we compared some Fusion methods horizontally, such as DenseASPP[42], ASPP[43] and SAPF (Scale-aware Pyramid Fusion)[44]. We replaced the Fusion module with different Fusion methods. Other modules remain unchanged, and the experimental results are shown in Table 3 benefit by the multiple pyramid module designed by us, the receptive field is not only

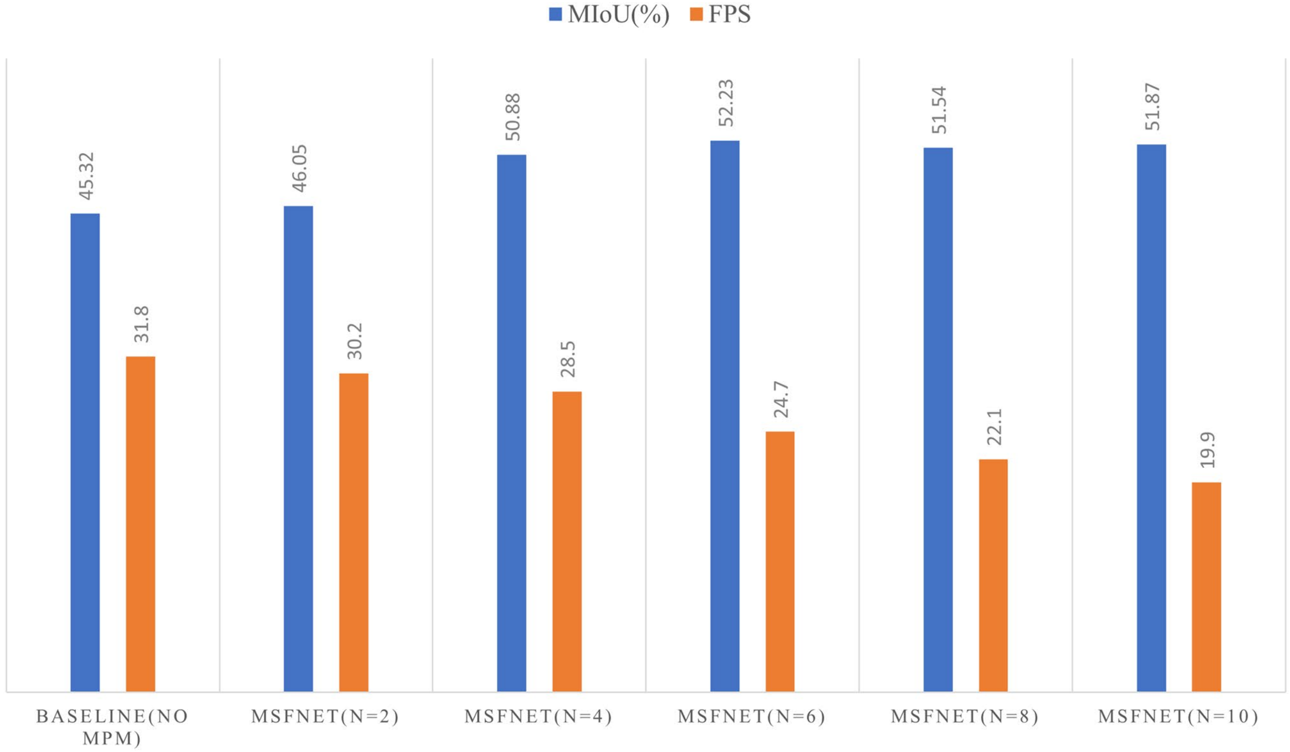

Figure 8. Result of segmentation.

Method	Params/M	MIoU (%)	FPS
Baseline + DenseASPP	76.25	48.86	15.3
Baseline + ASPP	44.35	48.80	22.5
Baseline + SAPF	71.18	47.22	18.3
MSFNet	48.64	52.23	24.7

Table 3. Ablation experiment of context module.

Model	WT-fusion	NCM	MPM	PA (%)	MPA (%)	MIoU (%)
Model A				73.9	58.8	45.3
Model B	√			74.6	61.0	49.48
Model C	√	√		75.9	61.3	49.98
Model D	√		√	76.3	62.8	51.66
WTNet	√	√	√	77.8	64.2	52.23

Table 4. Ablation Experiments.

increased in a single dimension, but also the feature information connection between contexts is strengthened through the splitting and stitching of channels, so as to improve the segmentation accuracy.

Ablation experiments of each module. This section analyzes the influence of the added wavelet transform modules (wavelet fusion module, wavelet transform replacement pooling module and wavelet transform connection module) on the semantic segmentation results. In order to analyze the influence of different modules, we constructed the four kinds of models, are respectively the ABCD, in A model, we don't add any module structure of encoder and decoder, fusion form is directly convolution operation, we only add in the model B wavelet fusion module, in C model, we added on the basis of the model B replace DWT pooling operation module, In model D, we add a context connection module on the basis of model B. Table 4 summarizes the experimental results of each model. Can be seen from the results of B add wavelet transform fusion module can effectively improve the performance of the semantic segmentation can be seen from the CD in the WTC module for network more

	PA (%)		MPA (%)		MIoU (%)	
Model	**NYUv2**	**SUN- RGBD**	**NYUv2**	**SUN- RGBD**	**NYUv2**	**SUN- RGBD**
RDF (2017)[45]	76.0	81.5	62.8	60.1	50.1	47.7
ACNet (2019)[27]	–	–	–	–	48.3	48.1
CTNet (2019)[46]	76.3	82.4	_	–	50.6	48.5
TSNet (2020)[47]	73.5	_	59.6	–	46.1	_
SGNet (2021)[30]	76.1	82.0	62.7	–	50.7	48.6
ESANet (2021)[37]	–	–	–	–	51.58	48.31
EMSANet (2022)[48]	–	–	–	–	53.34	48.47
FRNet (2022)[49]	77.6	87.4	66.5	62.2	53.6	51.8
MSFNet (ours)	**77.8**	**83.6**	**64.2**	**62.4**	**52.23**	**50.32**

Table 5. Results on Datasets.

Model	Backbone	MIoU (%)	Param (M)	FPS
RDF (2017)[45]	2ResNet101	49.1	169.1	11
ACNet (2019)[27]	2ResNet50	48.3	116.6	18
SGNet (2021)[30]	2ResNet50	47.7	39.3	39
ESANet (2021)[37]	2ResNet50	50.53	54.46	22.6
EMSANet (2022)[48]	2ResNet34	53.34	–	24.5
MSFNet (ours)	2ResNet50	52.23	48.64	24.7

Table 6. Inference Speed test on NYUv2 Dataset.

obvious indicators of ascension, this is because the wavelet transform for the outline of encoder and decoder joint can retain more details information, such as improving the precision of semantic segmentation. Compared with Model B, adding NCM module and MPM module improves the segmentation accuracy of model B by 0.5% and 2.2% respectively. The NCM module retains the frequency and orientation features of the image, and the MPM module plays the role of preserving the context information at the encoder-decoder junction transition stage, and fuses multi-scale features.

Results on NYUv2 and SUN-RGBD datasets. We tested our network model on NYUv2 and SUNRGBD datasets and compared it with existing network algorithms. The experimental results are shown in Table 5. The experimental results are shown in Table 6. Experimental results show that the proposed algorithm improves the pixel accuracy, the average pixel accuracy and the average intersection point ratio. The specific values on NYUv2 dataset are 77.8, 64.2 and 52.23%, and those on SUN-RGBD dataset are 83.6, 62.4 and 50.32%. Compared with the current SOTA algorithm, the three evaluation indexes have been improved. We believe that this is due to the combination of the wavelet transform RGB-D fusion, which retains more details and makes the two have better fusion effect. While at the same time only using the ResNet-50 coding structure, instead of using a network of deeper encoding but by replacing pooling layer as well as the design of the encoder and decoder connection module, wavelet transform to the edge of the target object information more accurate, at the same time, make the whole network is not deep structure of the encoder can get very good segmentation effect, the experimental results are shown in Table 5.

Experimental results show that our MSFNet has good performance on different datasets, which indicates that it can adapt to multiple categories and scenarios. Meanwhile, we test the parameter size and inference speed of MSFNet on NYUv2 dataset. The experiment was deployed on an NVIDIA 1080Ti. The results are shown in Table 6. It can be seen from the results that the wavelet transform module we designed only needs a small amount of extra computation, and still can achieve the speed of real-time inference.

At the same time, we compare the semantic segmentation results with RedNet, ACNet and ESANet[37] networks. The residual module in RedNet is used as a basic building block in encoder and decoder to construct a fusion structure and propose a pyramid supervised training scheme. ACNet uses independent branches based on ResNet to fuse RGB features and depth features, and finally obtains segmentation results after multiple upsampling. ESANet also adopts ResNet as the backbone network, adopts direct additive fusion in the fusion part, and adds a module similar to pyramid pooling at the encoder-decoder connection[50]. The segmentation results of each network are shown in Fig. 9, from which it can be seen that more details are preserved through the wavelet transform module. The segmentation results not only guarantee the contour details of the larger object, but also retain the information of some small object. The segmentation details are shown in Fig. 10.

From the segmentation details, it can be seen that thanks to the wavelet fusion and other modules designed by us, the edge contour information is relatively complete. For example, the wall and door frame details in the first line are smoother, and the edge contour details of the wash basin in the second line are also more abundant.

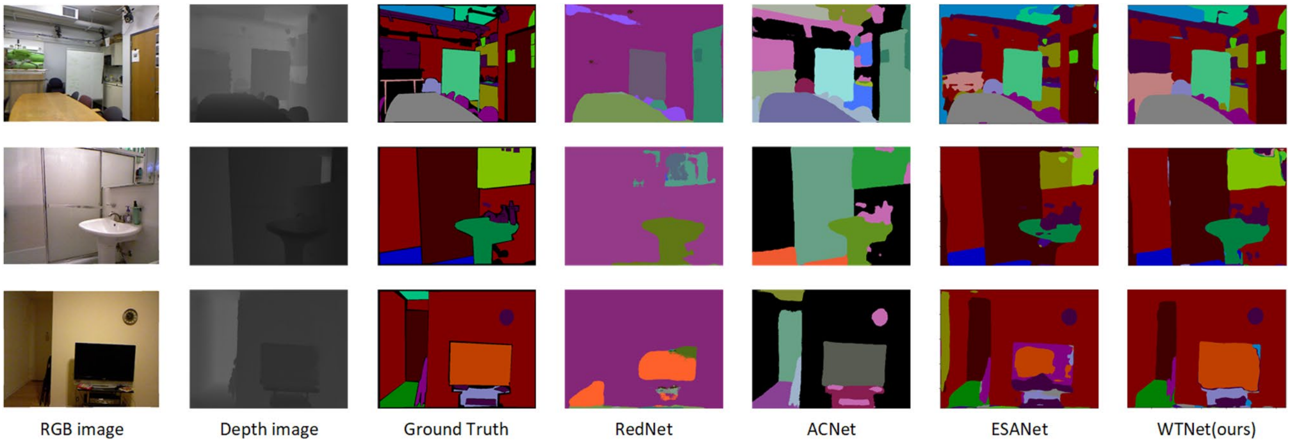

Figure 9. Result of segmentation.

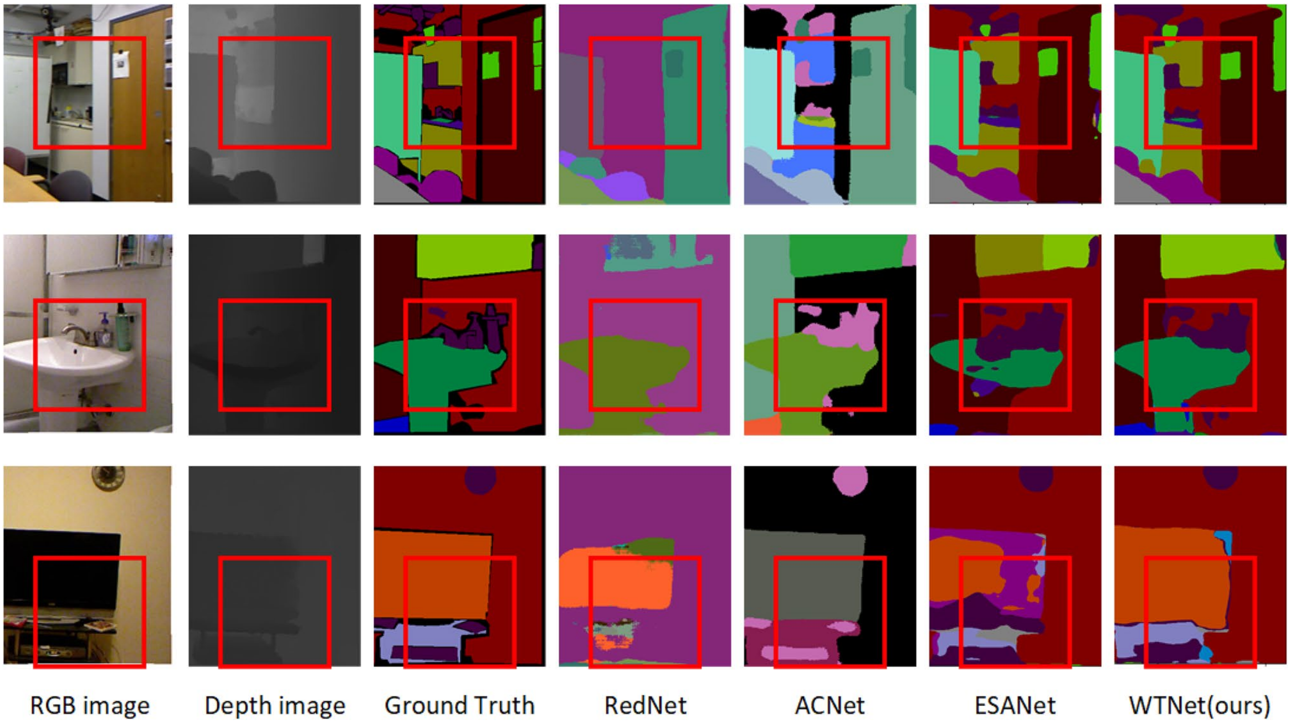

Figure 10. Detail of Segmentation result.

In the third row, although the TV edge is not as good as ACNet segmentation, the details in the TV cabinet are more complete.

Conclusion

In this study, we propose an RGB-D indoor semantic segmentation network based on multi scale fusion. Through the ResNet-50 as backbone of color image and the depth image encoder, add wavelet transform fusion module, and replace the pooling DWT wavelet module operation, and design a multiple pyramid module to aggregate multi-scale information and context global information, combined with low frequency and high frequency information fully, retains the outline details of the object, obtain more accurate information, and use nonsubsampled contour wave transform replaces the pooling operation. To further optimize the network segmentation effect, the encoder-decoder is connected by hop connection, and three side output supervised loss functions are added simultaneously. Experimental results on two datasets show that the performance of the proposed MSFNet is better than the current models, and the speed also meets the requirements of real-time inference, which provides a new method and idea for RGB-D indoor scene segmentation.

Data availability

The datasets generated and analysed during the current study are available in the NYUv2 repository and SUN-RGB-D repository, [https://cs.nyu.edu/silberman/datasets/nyu_depth_v2.html] and [https://rgbd.cs.princeton.edu/].

Received: 16 August 2022; Accepted: 21 November 2022
Published online: 24 November 2022

References

1. Kaut, H. & Singh, R. A review on image segmentation techniques for future research study. *Int. J. Eng. Trends Technol.* **35**, 504–505 (2016).
2. Dong, G., Yan, Y., Shen, C. & Wang, H. Real-time high-performance semantic image segmentation of urban street scenes. *IEEE Trans. Intell. Transp. Syst.* **22**, 3258–3274 (2020).
3. Yuan, X., Shi, J. & Gu, L. A review of deep learning methods for semantic segmentation of remote sensing imagery. *Expert Syst. Appl.* **169**, 114417 (2021).
4. Khan, M. Z., Gajendran, M. K., Lee, Y. & Khan, M. A. Deep neural architectures for medical image semantic segmentation. *IEEE Access* **9**, 83002–83024 (2021).
5. Hu, Y., Chen, Z. & Lin, W. Rgb-d semantic segmentation: A review. In *2018 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)* 1–6 (IEEE, 2018).
6. Couprie, C., Farabet, C., Najman, L. & LeCun, Y. Indoor semantic segmentation using depth information. arXiv preprint arXiv:1301.3572 (2013).
7. Jiang, J., Zheng, L., Luo, F. & Zhang, Z. Rednet: Residual encoder-decoder network for indoor rgb-d semantic segmentation. arXiv preprint arXiv:1806.01054 (2018).
8. Li, Y., Zhang, J., Cheng, Y., Huang, K. & Tan, T. Semantics-guided multi-level rgb-d feature fusion for indoor semantic segmentation. In *2017 IEEE International Conference on Image Processing (ICIP)* 1262–1266 (IEEE, 2017).
9. Chang, M., Guo, F. & Ji, R. Depth-assisted refinenet for indoor semantic segmentation. In *2018 24th International Conference on Pattern Recognition (ICPR)* 1845–1850 (IEEE, 2018).
10. Li, Q., Shen, L., Guo, S. & Lai, Z. Wavecnet: Wavelet integrated cnns to suppress aliasing effect for noise-robust image classification. *IEEE Trans. Image Process.* **30**, 7074–7089 (2021).
11. Mallat, S. G. A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Trans. Pattern Anal. Mach. Intell.* **11**, 674–693 (1989).
12. Liu, P., Zhang, H., Lian, W. & Zuo, W. Multi-level wavelet convolutional neural networks. *IEEE Access* **7**, 74973–74985 (2019).
13. Ramamonjisoa, M., Firman, M., Watson, J., Lepetit, V. & Turmukhambetov, D. Single image depth estimation using wavelet decomposition. arXiv preprint arXiv:2106.02022 (2021).
14. Xu, K. *et al.* Learning in the frequency domain. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 1740–1749 (2020).
15. Silberman, N., Hoiem, D., Kohli, P. & Fergus, R. Indoor segmentation and support inference from rgbd images. In *European Conference on Computer Vision* 746–760 (Springer, 2012).
16. Song, S., Lichtenberg, S. P. & Xiao, J. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 567–576 (2015).
17. Long, J., Shelhamer, E. & Darrell, T. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 3431–3440 (2015).
18. Ronneberger, O., Fischer, P. & Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* 234–241 (Springer, 2015).
19. Badrinarayanan, V., Kendall, A. & Cipolla, R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **39**, 2481–2495 (2017).
20. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 770–778 (2016).
21. Zhao, H., Shi, J., Qi, X., Wang, X. & Jia, J. Pyramid scene parsing network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 2881–2890 (2017).
22. Borse, S., Cai, H., Zhang, Y. & Porikli, F. Hs3: Learning with proper task complexity in hierarchically supervised semantic segmentation. arXiv preprint arXiv:2111.02333 (2021).
23. Li, Y. *et al.* X-net: A dual encoding–decoding method in medical image segmentation. *Vis. Comput.* 1–11 (2021).
24. He, Y., Chiu, W.-C., Keuper, M. & Fritz, M. Std2p: Rgbd semantic segmentation using spatio-temporal data-driven pooling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 4837–4846 (2017).
25. Eigen, D. & Fergus, R. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *Proceedings of the IEEE International Conference on Computer Vision* 2650–2658 (2015).
26. Hazirbas, C., Ma, L., Domokos, C. & Cremers, D. Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture. In *Asian Conference on Computer Vision* 213–228 (Springer, 2016).
27. Hu, X., Yang, K., Fei, L. & Wang, K. Acnet: Attention based network to exploit complementary features for rgbd semantic segmentation. In *2019 IEEE International Conference on Image Processing (ICIP)* 1440–1444 (IEEE, 2019).
28. Gupta, S., Girshick, R., Arbeláez, P. & Malik, J. Learning rich features from rgb-d images for object detection and segmentation. In *European Conference on Computer Vision* 345–360 (Springer, 2014).
29. Xing, Y., Wang, J., Chen, X. & Zeng, G. Coupling two-stream rgb-d semantic segmentation network by idempotent mappings. In *2019 IEEE International Conference on Image Processing (ICIP)* 1850–1854 (IEEE, 2019).
30. Chen, L.-Z., Lin, Z., Wang, Z., Yang, Y.-L. & Cheng, M.-M. Spatial information guided convolution for real-time rgbd semantic segmentation. *IEEE Trans. Image Process.* **30**, 2313–2324 (2021).
31. Chen, S., Zhu, X., Liu, W., He, X. & Liu, J. Global-local propagation network for rgb-d semantic segmentation. arXiv preprint arXiv:2101.10801 (2021).
32. Cao, J. *et al.* Shapeconv: Shape-aware convolutional layer for indoor rgb-d semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* 7088–7097 (2021).
33. Bae, W., Yoo, J. & Chul Ye, J. Beyond deep residual learning for image restoration: Persistent homology-guided manifold simplification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* 145–153 (2017).
34. Guo, T., Seyed Mousavi, H., Huu Vu, T. & Monga, V. Deep wavelet prediction for image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* 104–113 (2017).
35. Li, Q., Shen, L., Guo, S. & Lai, Z. Wavelet integrated cnns for noise-robust image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 7245–7254 (2020).
36. Li, Y., Wang, Y., Leng, T. & Zhijie, W. Wavelet u-net for medical image segmentation. In *International Conference on Artificial Neural Networks* 800–810 (Springer, 2020).

37. Seichter, D., Köhler, M., Lewandowski, B., Wengefeld, T. & Gross, H.-M. Efficient rgb-d semantic segmentation for indoor scene analysis. In *2021 IEEE International Conference on Robotics and Automation (ICRA)* 13525–13531 (IEEE, 2021).
38. Mallat, S. G. A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Trans. Pattern Anal. Mach. Intell.* **11**, 674–693 (1989).
39. Hu, J., Shen, L. & Sun, G. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 7132–7141 (2018).
40. Howard, A. G. *et al.* Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861 (2017).
41. Pan, S. J. & Yang, Q. A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* **22**, 1345–1359 (2009).
42. Yang, M., Yu, K., Zhang, C., Li, Z. & Yang, K. Denseaspp for semantic segmentation in street scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 3684–3692 (2018).
43. Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K. & Yuille, A. L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.* **40**, 834–848 (2017).
44. Feng, S. *et al.* Cpfnet: Context pyramid fusion network for medical image segmentation. *IEEE Trans. Med. Imaging* **39**, 3008–3018 (2020).
45. Park, S.-J., Hong, K.-S. & Lee, S. Rdfnet: Rgb-d multi-level residual feature fusion for indoor semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision* 4980–4989 (2017).
46. Xing, Y., Wang, J., Chen, X. & Zeng, G. Coupling two-stream rgb-d semantic segmentation network by idempotent mappings. In *2019 IEEE International Conference on Image Processing (ICIP)* 1850–1854 (IEEE, 2019).
47. Zhou, W., Yuan, J., Lei, J. & Luo, T. Tsnet: Three-stream self-attention network for rgb-d indoor semantic segmentation. *IEEE Intell. Syst.* **36**, 73–78 (2020).
48. Seichter, D., Fischedick, S. B., Köhler, M. & Groß, H.-M. Efficient multi-task rgb-d scene analysis for indoor environments. In *2022 International Joint Conference on Neural Networks (IJCNN)* 1–10 (IEEE, 2022).
49. Zhou, W., Yang, E., Lei, J. & Yu, L. Frnet: Feature reconstruction network for rgb-d indoor scene parsing. *IEEE J. Sel. Top. Signal Process.* (2022).
50. Zhao, H., Shi, J., Qi, X., Wang, X. & Jia, J. Pyramid scene parsing network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 2881–2890 (2017).

Acknowledgements

This work was supported by Guizhou Science and Technology Planning Project (Guizhou Science and Technology Cooperation Support [2021] General 176).

Author contributions

S.J. and D.L. and R.F. conceived the experiments, S.J. and R.F. conducted the experiment(s), S.J. and Y.X. analysed the results. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Y.X.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.