Application of Multi-modal Fusion Attention Mechanism in Semantic Segmentation

Yunlong Liu(✉)[0000−0003−0545−7985], Osamu Yoshie[0000−0002−4192−554X], and Hiroshi Watanabe[0000−0002−9306−688X]

Faculty of Science and Engineering, Waseda University, Japan
liuyunlong@akane.waseda.jp
{yoshie, hiroshi.watanabe}@waseda.jp

Abstract. The difficulty of semantic segmentation in computer vision has been reintroduced as a topic of interest for researchers thanks to the advancement of deep learning algorithms. This research aims into the logic of multi-modal semantic segmentation on images with two different modalities of RGB and Depth, which employs RGB-D images as input. For cross-modal calibration and fusion, this research presents a novel FFCA Module. It can achieve the goal of enhancing segmentation results by acquiring complementing information from several modalities. This module is plug-and-play compatible and can be used with existing neural networks. A multi-modal semantic segmentation network named FFCANet has been designed to test the validity, with a dual-branch encoder structure and a global context module developed using the classic combination of ResNet and DeepLabV3+ backbone. Compared with the baseline, the model used in this research has drastically improved the accuracy of the semantic segmentation task.

1 Introduction

Semantic segmentation has been a fundamental and critical problem in computer vision [1] for the past long time. The advent of neural networks significantly improved semantic segmentation performance [2]. CNN based on the Encoder-Decoder structures [3] has become the mainstream, and its accuracy and efficiency have greatly exceeded other methods.

However, semantic segmentation in indoor scenes[4] is still a challenging task, because its semantic information is more complex than outdoor scene[5]. Multi-modal input is a feasible solution for indoor scene [6]. The RGB-D method is becoming increasingly popular among multi-modal segmentation systems since it may gather spatial information and scene structure coding.

Previously, the biggest obstacle to RGB-D solutions was that depth images were difficult to obtain. However, with the proliferation of depth cameras[7], collecting depth photos is no longer expensive and complicated. The popularity of civilian depth cameras [8], such as the Microsoft Kinect sensor [6], has made acquiring depth photos much easier and cheaper. Consequently, semantic segmentation datasets based on RGB-D images have started to arise sequentially

[9]. RGB-D-based semantic segmentation is becoming more and more popular and common.

Further modalities bring additional spatial information, but fusing these two modalities becomes new challenge [10]. Calculating HHA [11] is a reasonable way to enhance depth in preprocessing phase for extracting features better. HHA includes three channels: horizontal disparity, height above ground, and the angle the pixel's local surface normal makes with the inferred gravity direction [12].

In the deep-learning-based RGB-D segmentation methods, to extract features from RGB and depth modalities separately, it is common to design a dual branching network structure [13]. The depth features are gradually fused into the feature map of RGB when the network goes deeper. Adding features with depth information to RGB can significantly improve the accuracy of segmentation [14]. Especially in the case of using HHA, since the HHA image has three channels, the same structure can be used to extract RGB and depth information.

In this paper, We innovatively propose a multi-modal feature fusion attention mechanism for cross-modal calibration, which calibrates the information of different modalities in both feature channels and spatial dimensions through two different attention mechanisms. The unique feature fusion module enables the network to better capture the complementary information among different modalities based on the standard features, thus improving the segmentation results.

According to its characteristics, this module is named the feature fusion cascade attention mechanism module, abbreviated as the FFCA module. The method chapter will introduce this module, including multi-modal attention and feature fusion, and show the semantic segmentation model based on this FFCA module: FFCANet.

2 Related Work

2.1 Semantic Segmentation

In computer vision, the first application of CNN is image classification [15]. Consequently, other directions also emerged with deep learning-based approaches, such as detection, generation, and segmentation. Similar to the concept of "encoding" in NLP [16], the Neural Network for semantic segmentation task usually consists of two parts: The part that extracts the feature map is called the encoder, and the part that obtains the segmentation result from the feature map is called the decoder.

Encoder is used to extract the feature map though cropping or compressing the input image, they generally have similar structure with the backbone used for classification task [17], such as AlexNet [18] and so on [19] [20] [21] [22] [23]. These networks focus on how to extract features from the input image more efficiently, then have gradually developed several different styles. Some networks choose to utilize convolutional kernels of different sizes to obtain multi-scale contextual features [19] [24] [25] [26]; others desire to make more efficient use of the

information available in the image while reducing the computational overhead DenseNet [27] [28] [29]. The most popular network is the ResNet families [21]. It adds identity connections to solve the degradation problem in deep models. This scheme the most common encoder for semantic segmentation tasks.

The word decoder first appeared in SegNet [3] through a series of deconvolution [30] or up-sampling operations. The decoder restores the feature map to the original input size to realize the pixel-by-pixel classification of the input image. Compared with the traditional CNN structure, these decoders generally have a different convolutional structure [31] [32] [33]. These kernels use multiple sizes of the kernel to capture features at different scales to solve the multi-scale problem [34] [35] [36]. A famous example is DeepLab series [37] [38] [39] [40] . This structure is widely used in various tasks related to semantic segmentation. The most advanced DeepLabv3+ [40] in the series employs atrous convolution and spatial pyramidal pooling module and simultaneously improves speed and accuracy.

Another unique structure is the global context module. It can incorporate global information into the feature map. Before the emergence of CNN, context modules are already present in the traditional approaches [41] [42]. Subsequently, this concept was introduced to deep learning [43] [44] [45] [46]. A more general approach is to insert the context module between the encoder and the decoder. The most typical example of this structure is EncNet [47]. By introducing the context encoding module, EncNet is capable of processing global contextual information.

2.2 RGB-D Segmentation

As a kind of multi-modal approach [48], using RGB-D images as input are particularly common in the segmentation task of indoor scenes. Since RGB-D photographs contain spatial information missing from 2D images, they can reveal more relevant elements in indoor situations unrelated to lighting. Couprie et al. [14] found that using depth information can significantly improve the distinguishability of objects with a similar appearance.

Researchers discovered the effectiveness of deep information quite early, even before deep learning emerged. The traditional method of RGB-D segmentation can be considered as a relatively fixed pipeline mode [49]. Similar to traditional 2D image segmentation, pre-segmentation segments the original RGB-D image into more basic units, such as superpixels [50], blocks [51] and regions [52], then these units will be classified by traditional statistical [53] [54] [55] or machine learning methods [56] [12]. The earliest attempt came from Silberman et al [5]. They created the NYUv1 data set, which was eventually expanded to NYUv2 [4]. This approach based on Conditional Random Field (CRF) and has inspired many subsequent researches [57] [58] [59], even in some early models of deep learning also containing CRF-like structures[60] [37] [38].

The subsequent advent of deep learning likewise brought significant developments to RGB-D segmentation. These neural networks usually use two encoder branches to extract RGB and depth features separately. This structure was first

seen in FuseNet [13]. The focus of this type of research is how to integrate RGB and depth information. Common strategies include Early Fusion [14], Middle Fusion [30] or Late Fusion [61]. These early approaches inspired later research, such as RDFNet [62] which extends RefineNet [63] with a multi-modal fusion block. LSTM-CF [64] uses a horizontal LSTM to capture RGB and depth features, then introduces a vertical LSTM to combine them. Liu et al. improved the HHA [11] encoding by integrating 2D and 3D information [61]. They also extended the VGG [14] encoder in DeepLab [37] for RGB-D semantic segmentation. Cheng et al. proposed a gated fusion method [30], which studied the paired relationship between adjacent RGB-D pixels. MTI-Net [65] discussed the importance of considering task interaction on multiple scales when extracting task information in a multi-task learning setting. ICM [66] is an ensemble classification model that proposes regularization based on variance. ESANet [67] has an encoder with two branches and uses the attention mechanism to fuse depth into the RGB encoder in several stages. ShapeConv [68] introduces a shape-aware convolution layer to process depth features.

The method used in this paper is to combine the mid-term fusion of the attention mechanism, use the attention mechanism to calibrate the features from different modalities, and fuse the features of each stage of the encoder through element-by-element addition.

3 Method

After a detailed discussion of the strengths and weaknesses of best practices for RGB-D semantic segmentation, we propose a new method for RGB-D semantic segmentation, inspired by attention mechanisms like CBAM [69] and SKNet [70]. The crucial part of this method is a Feature Fused Cascade Attention Module (FFCA Module).

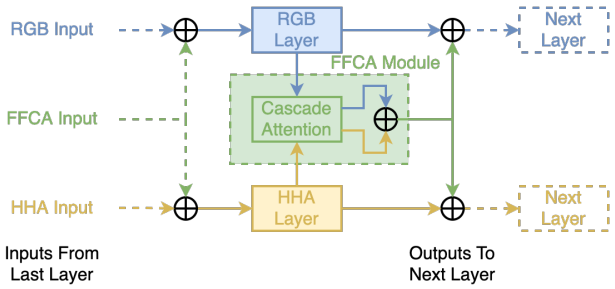

Fig. 1. FFCA module

Considering that each down-sampling occurs at the beginning of the hidden layers, the output of each hidden layer is the feature map with the highest level of

abstraction. Therefore, feature fusion should occur after each hidden layer. Fig. 1 shows the insertion method and internal structure of the FFCA Module proposed in this study. Different features from two different modalities are converged into the RGB and HHA branches in an element-by-element summation after a cross-modal calibration based on the Attention mechanism to achieve the fusion of features. HHA [11] is a generalization of Depth images that makes it easier to apply CNN algorithms to RGB-D data, which is a considerable improvement over depth channels alone.

3.1 FFCA Module: Feature Fused Cascade Attention Module

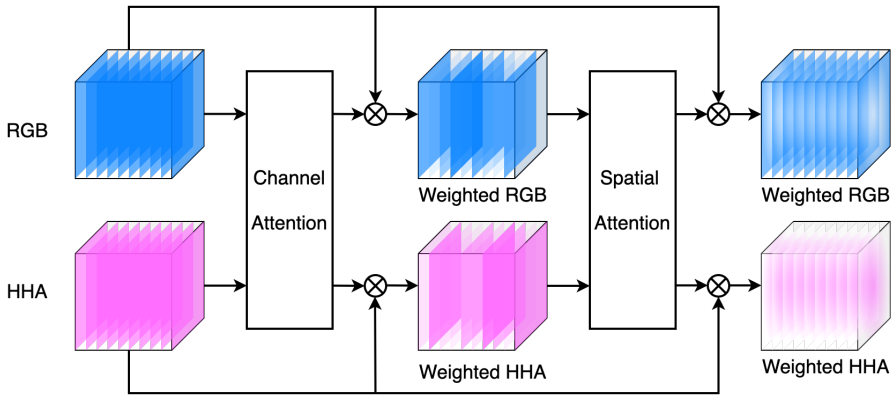

Fig. 2. Cascaded Channel and Spatial Attention

The extra spatial information contained in RGB-D can compensate well for the lack of RGB compared to pure RGB's traditional 2D image semantic segmentation. However, simply adding the output features from the two coding branches may not achieve the desired result due to the difficulty of aligning the depth information with RGB and the amount of noise it contains. Therefore, the critical of feature fusion is to handle the differences between two different image signals properly.

This research proposes a cross-modal Cascade Attention to solve these problems. As shown in Fig. 2, this structure contains two different attention mechanisms: Channel Attention and Spatial Attention. This module concatenates the two attention structures to perform cross-modal calibration in the feature maps' channel and spatial dimensions. The feature maps of both modalities have the same size. The calibration assigns a pair of weights for elements at the same position in two modalities to facilitate subsequent feature fusion by element-by-element addition.

3.2 Multiple Layer Channel Attention

The structure of multiple layer channel attention in cascaded attention shown in Fig. 3. This multilayer Channel Attention can be regarded as a modified solution of the multi-input attention mechanism in SKNet [70]. The difference is that the two feature maps come from two different modalities.

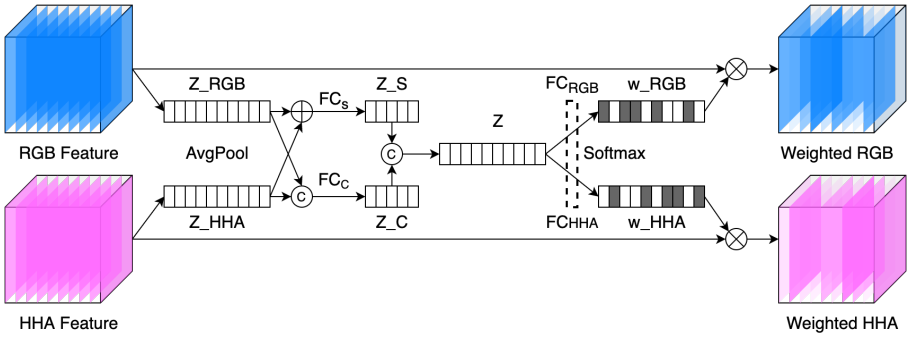

Fig. 3. Multiple Layer Channel Attention

This channel attention has two layers: the first concatenates and sums the feature vectors, thus extracting the separated and fused features of the two modalities. The second concatenates these two features, then uses two fully connected layers and softmax to obtain the RGB and HHA attention vectors. Taking any RGB feature map $RGB \in \mathbb{R}^{H\times W\times C}$, the HHA feature map $HHA \in \mathbb{R}^{H\times W\times C}$ as input, the operation of this Channel Attention can be described in the following mathematical language:

Global averaging pooling: like the Squeeze operation in SENet [71], the RGB and HHA feature maps need to undergo a global averaging pooling $\mathcal{F}_{gp}$ after being fed into the Attention module to obtain the feature vectors $z_{RGB} \in \mathbb{R}^{C}$ and $z_{HHA} \in \mathbb{R}^{C}$, enabling them to be fed into the subsequent fully connected layer. Specifically, the c-th element of both vectors is computed utilizing the c-th channel of shape H × W in the corresponding feature map:

$$z_{RGB_c} = \mathcal{F}_{gp}\left(RGB_c\right) = \frac{1}{H \times W}\sum_{i=1}^{H}\sum_{j=1}^{w} RGB_c(i,j) \tag{1}$$

$$z_{HHA_c} = \mathcal{F}_{gp}\left(HHA_c\right) = \frac{1}{H \times W}\sum_{i=1}^{H}\sum_{j=1}^{w} HHA_c(i,j) \tag{2}$$

Multi-layered full-connection: the role of the two full-connection layers in the Attention block is to produce a Scale vector z that fuses two different

modalities. to be able to take advantage of the common features of both modalities while calibrating across modalities, the inputs to the first full-connection layer are $\mathrm{z_{RGB}} \in \mathbb{R}^{\mathrm{C}}$ and $\mathrm{z_{HHA}} \in \mathbb{R}^{\mathrm{C}}$ summing element by element the fusion vector at $\mathrm{z_{sum}} \in \mathbb{R}^{\mathrm{C}}$, and $\mathrm{z_{concat}} \in \mathbb{R}^{2\mathrm{C}}$ obtained by concatenating the two. The first layer extracts the fusion feature $\mathrm{z_s} \in \mathbb{R}^{\mathrm{d}}$ and the separation feature $\mathrm{z_c} \in \mathbb{R}^{\mathrm{d}}$ for each of the two modalities through two fully connected layers $\mathcal{F}_{\mathcal{FC}_{\mathrm{s}}}$ and $\mathcal{F}_{\mathcal{FC}_{\mathrm{c}}}$ which are at the same level:

$$\mathrm{z_s} = \mathcal{F}_{\mathcal{FC}_{\mathrm{s}}}\left(\mathrm{z_{sum}}\right) = \delta\left(\mathcal{B}\left(\mathrm{W_s} \times \left(\mathrm{z_{RGB}} \oplus \mathrm{z_{HHA}}\right)\right)\right) \tag{3}$$

$$z_s = \mathcal{F}_{\mathcal{FC}_s}\left(z_{\mathrm{sum}}\right) = \delta\left(\mathcal{B}\left(W_s \times \left(z_{RGB} \| z_{HHA}\right)\right)\right) \tag{4}$$

Where δ represents the ReLU activation layer, $\mathcal{B}$ is the Batch Norm layer, and $\mathrm{W_s} \in \mathbb{R}^{\mathrm{d}\times\mathrm{c}}$ and $\mathrm{W_c} \in \mathbb{R}^{\mathrm{d}\times 2\mathrm{C}}$ are the weight parameters for the fully connected layers $\mathcal{F}_{\mathcal{FC}_s}$ and $\mathcal{F}_{\mathcal{FC}_{\mathrm{c}}}$ weight parameters. d is the number of channels reduced after a fully connected squeeze and the minimum value is given via L. The shrinkage ratio r and the minimum value L are both hyperparameters of the network structure. Typically, $r = 16$ and $L = 32$. This value is dynamically adjusted by itself in a similar way to that in SENet and proportion to a certain range:

$$\mathrm{d} = \max(\mathrm{C}/\mathrm{r}, \mathrm{L}) \tag{5}$$

The second layer of full concatenation is equivalent to the excitation operation in SENet and is used to obtain the weights of both RGB and HHA modalities in the channel dimension. The fused features z_s and separated features z_c in the first layer are stitched together into a feature vector $z \in \mathbb{R}^{2\mathrm{d}}$ of twice the length, which contains both fused and separated features for both RGB and HHA modalities, making the fully connected layer $\mathcal{F}_{\mathcal{FC}_{\mathrm{RGB}}}$, which is located in the second layer for the two different modalities, and $\mathcal{F}_{\mathcal{FC}_{\mathrm{HHA}}}$ can extract the required feature weights for each. Similar to the first layer, the weights of the two full connections are $\mathrm{W_{RGB}} \in \mathbb{R}^{\mathrm{C}\times 2\mathrm{d}}$ and $\mathrm{W_{HHA}} \in \mathbb{R}^{\mathrm{C}\times 2\mathrm{d}}$ respectively.

$$w_{RGB} = \mathcal{F}_{\mathcal{FC}_{RGB}}(z) = W_{RGB} \times \left(z_S \| z_c\right) \tag{6}$$

$$w_{HHA} = \mathcal{F}_{\mathcal{FC}_{HHA}}(z) = W_{HHA} \times \left(z_s \| z_c\right) \tag{7}$$

Cross-modal Softmax normalization: multi-layer full connectivity has filtered out those feature channels from RGB and HHA that are more useful for subsequent segmentation tasks and given them higher weights. However, some of the corresponding channels of the two feature maps may be redundant or contain some information that would interfere with each other. For the subsequent feature fusion to proceed smoothly, using softmax to calibrate the feature weights jointly w_{RGB} and w_{HHA} is necessary.

$$w_{RGB}(\mathrm{Calibrated}) = \mathcal{F}_{\mathrm{Softmax}}\left(w_{RGB} \| w_{HHA}\right)_{\mathrm{dim}=1}\left[w_{RGB}\right] \tag{8}$$

$$w_{RGB}(\mathrm{Calibrated}) = \mathcal{F}_{\mathrm{Softmax}}\left(w_{RGB} \| w_{HHA}\right)_{\mathrm{dim}=1}\left[w_{HHA}\right] \tag{9}$$

In this process, the weight vectors w_{RGB} and w_{HHA} are no longer spliced into longer vectors, but adding a matrix of dimension $W_{RGB||HHA} \in R^{2\times C}$, the normalization of Softmax is performed in this new extended dimension. This normalization allows the weights of the feature channels at the corresponding positions of the two modalities to always sum to 1, enabling the maximum exploitation of the complementary features of the different modalities on the channels.

3.3 Fusion Spatial Attention

This research addresses RGB-D feature fusion problems by introducing a Fusion Spatial Attention module. Compared with the general single-mode space Attention mechanism, the multi-modal space fusion attention includes inputs from two different modalities in the final convolution process. Fig. 4 shows this structure. In addition to splicing the pooled single-channel features, the spatial feature also contains an additional mixing channel.

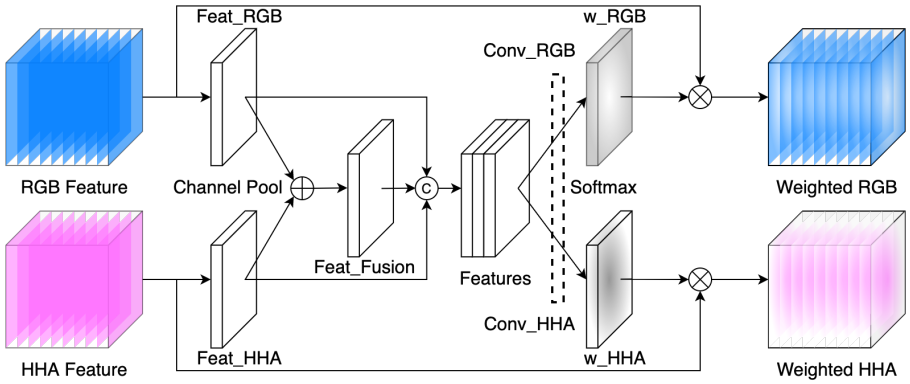

Fig. 4. Fusion Spatial Attention

Taking as input an arbitrary RGB feature map $RGB \in \mathbb{R}^{H\times W\times C}$ and an HHA feature map $HHA \in \mathbb{R}^{H\times W\times C}$, the following mathematical language will describe this Spatial Attention operation:

Channel averaging pooling: channel pooling $\mathcal{F}_{cp}$ is a compression of the feature map from the channel dimension. A feature map of dimension $\mathrm{H}\times\mathrm{W}\times\mathrm{C}$ will be compressed to $\mathrm{H}\times\mathrm{W}$, keeping only one channel. The value of pixel (i.j) at any position in this single-channel feature map is the mean value of the pixels at the corresponding position for all channels in the original feature map:

$$\text{feat}_{\text{RGB}}(\text{i},\text{j}) = \mathcal{F}_{cp}(\text{RGB}) = \frac{1}{\text{C}}\sum_{\text{c}=1}^{\text{C}}\text{RGB}_{\text{c}}(\text{i},\text{j}) \tag{10}$$

$$\text{feat}_{\text{HHA}}(\text{i},\text{j}) = \mathcal{F}_{cp}(\text{HHA}) = \frac{1}{\text{C}}\sum_{\text{c}=1}^{\text{C}}\text{HHA}_{\text{c}}(\text{i},\text{j}) \tag{11}$$

Fusion channels: Spatial attention generally contains two pooling operations, to obtain RGB and depth feature channel. The two feature channel are concatenated into a two-channel feature map to provide redundant information.

On this basis, a hybrid channel is also innovatively introduced in this study to obtain the fusion information between the two modalities. The spatial features from two modalities are summed pixel-by-pixel to let subsequent convolution exploit the complement information better:

$$\text{feat}_{\text{Fusion}} = \text{feat}_{\text{RGB}} \oplus \text{feat}_{\text{HHA}} \tag{12}$$

$$\text{feature} = \text{feat}_{\text{RGB}} \| \text{feat}_{\text{Fusion}} \| \text{feat}_{\text{HHA}} \tag{13}$$

Spatial weights: This Attention module uses two convolutional layers of the same size $\mathcal{F}_{Conv_{RGB}}$ and $\mathcal{F}_{Conv_{HHA}}$ to generate the Spatial Attention weights for RGB and HHA respectively. The Sigmoid activation function is discarded here and used for subsequent cross-modal calibration. The size of both sets of convolution kernels is 7×7, cause the larger size allows for the aggregation of more prodomain features:

$$w_{\text{RGB}} = \mathcal{F}_{Conv_{\text{RGB}}}\ (\text{features}) \ = \text{Conv}_{\text{RGB}}^{7\times7}\ (\text{features}) \tag{14}$$

$$w_{\text{HHA}} = \mathcal{F}_{C_{\sigma n \text{H}_{\text{HHA}}}}\ (\text{features}) \ = \text{Conv}_{\text{HHA}}^{7\text{x}7}\ (\text{features}) \tag{15}$$

Cross-modal Softmax normalization: Similar to Channel Attention in the previous section, Softmax is used here for joint spatial calibration, which will normalize the matrix of spatial attention. This is to address the signal alignment problem of RGB and HHA:

$$w_{RGB_{(\text{Calibrated})}} = \mathcal{F}_{\text{Softmax}}\left(w_{RGB} \| w_{HHA}\right)_{\text{dim}=1}\left[w_{RGB}\right] \tag{16}$$

$$w_{RGB_{(\text{Calibrated})}} = \mathcal{F}_{\text{Softmax}}\left(w_{RGB} \| w_{HHA}\right)_{\text{dim}=1}\left[w_{HHA}\right] \tag{17}$$

Since w_{RGB} and w_{HHA} are no longer feature vectors but one-dimensional feature maps, the stitched $w_{RGB\|HHA} \in R^{2\times W\times H}$ will have a three-dimensional shape. softmax normalization is still performed in this new extended dimension of length 2. This step allows the weights of the two sets of feature maps corresponding to spatial locations to always sum to 1, which can compensate for the different responses of the RGB and depth maps at the edges of the object and promote a better alignment of the two signals.

3.4 FFCANet: Feature Fused Cascade Attention Network

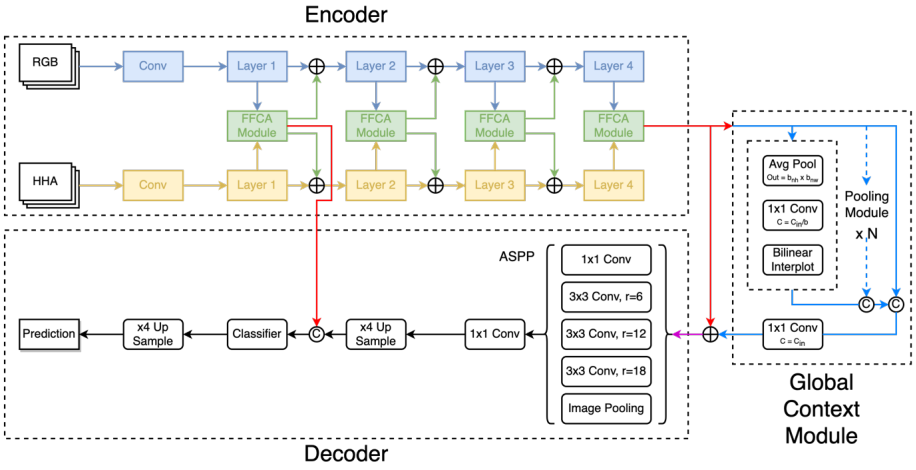

Fig. 5. Network structure of FFCANet

The FFCA Module is a plug-and-play cross-modal calibration and feature fusion module based on the Attention mechanism. Therefore, it also requires a network structure to host the module for the semantic segmentation task. We have built a network structure for the semantic segmentation task by modifying existing network components based on existing research. This network efficiently combines with the FFCA Module. We named it Feature Fused Cascade Attention Network, or FFCANet for short.

The structure of the network in this study shown in Fig. 5. The overall structure of the network consists of an encoder, context module, and decoder. The encoder part is chosen from ResNet [21], which is most commonly used in semantic segmentation tasks and is extended into two branches connected by FFCA Module. Context module is similar to pyramid pooling in PSPNet [34], refers to Seichter et al.'s scheme used in ESANet [72] and modified their approach. Since the two-branch structure of the encoder doubles the network parameters, DeepLabV3+ [40] with a smaller number of parameters was chosen for the decoder to balance the accuracy of the network with the memory overhead.

4 Experiment

To verify the validity of the innovative work made in this study, the NYUv2 [4] dataset was used as the benchmark for testing. Subsequent ablation experiments will also be conducted on this dataset.

4.1 Dataset and Metrics

Due to the scarcity of indoor RGBD datasets for semantic segmentation, NYU Depth v2 (NYUv2) has been the gold standard in this direction for the past few years. The dataset contains 1449 accurately labeled images with depth information, of which 795 are for the training set and 654 for the test set.

Semantic segmentation is an intensive classification task. It means each pixel in an image should be predicted to a semantic category. Therefore, we chose MIoU as this research's most dominant evaluation metric, like other semantic segmentation tasks. MIoU mean is Mean Intersection over Union. It is generally computed based on classes, and the IoU of each class is computed and then accumulated and averaged to obtain a global-based evaluation.

$$\mathrm{MIoU} = \frac{1}{\mathrm{k}+1} = \sum_{\mathrm{i}=0}^{\mathrm{k}} \frac{\mathrm{p_{ii}}}{\sum_{\mathrm{j}=0}^{\mathrm{k}} \mathrm{p_{ij}} + \sum_{\mathrm{j}=0}^{\mathrm{k}} \mathrm{p_{ji}} - \mathrm{p_{ii}}} \tag{18}$$

4.2 Optimizing Experience

This experiment aims to investigate the best way to use the optimal FFCA Module with the contextual module. We conduct experiments on three Backbone with two FFCA Module combining strategies: ResNet50, ResNet101, and ResNet152. The results are shown in Table 1.

Table 1. Result of optimize experience

Backbone	FFCAM	Context	P Acc	M Acc	FW Acc	MIoU
ResNet50	5	ppm-1357	76.80	62.01	63.57	50.11
	5	ppm-15	76.52	62.43	63.84	50.46
	4	ppm-1357	77.16	62.67	63.91	51.08
	4	ppm-15	76.96	62.71	63.90	51.19
ResNet101	5	ppm-1357	77.47	63.03	64.02	51.54
	5	ppm-15	77.78	62.98	64.22	51.81
	4	ppm-1357	77.92	63.28	64.77	52.32
	4	ppm-15	78.13	63.14	64.86	52.58
ResNet152	5	ppm-1357	77.32	63.87	64.89	52.53
	5	ppm-15	77.81	63.92	65.01	52.59
	4	ppm-1357	78.01	64.71	65.08	53.09
	4	ppm-15	78.39	65.31	65.72	53.30

We plug FFCA Module after each hidden layer for feature fusion. In addition to using four modules, another fusion strategy uses five modules, which the FFCA Module also inserted after the initial first convolutional layer. Two different resolution combinations have been experimented with for the contextual

modules to find the best combination of pyramidal pooling sizes. One was a two-way pooling branch of 1x1 and 5x3; the other was a four-way pooling branch of 1x1, 3x2, 5x3, and 7x5.

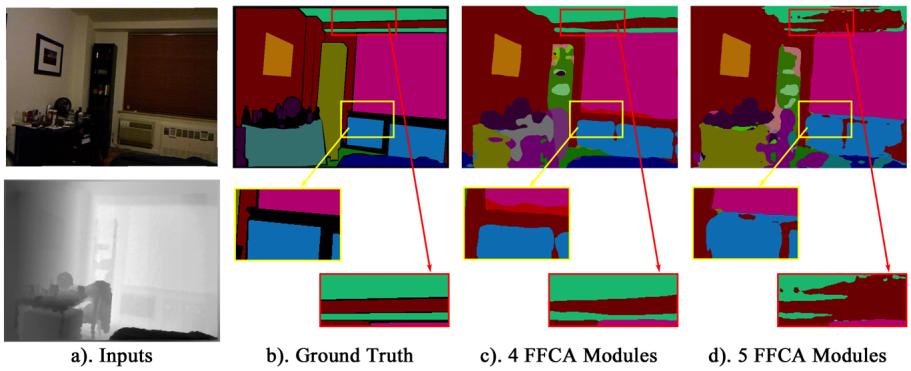

Fig. 6. Local edge detail of the result

The encoder structure using four FFCA Modules is optimal in the optimizing experiment, with the best calibration and fusion of features between different modes. Fig. 6 shows this visualized result. The network structure using 4 FFCA Module performs better at the edges of different objects, as shown in the yellow bordered area. They contain less inter-adhesion in the transition region. In contrast, at the locations marked by the red borders, the 5 FFCA Module structure classification results show many broken edge features, indicating that the RGB and depth signals are not well aligned.

Table 2. Comparison result

Method	P Acc	M Acc	MIoU
FCN [17]	65.4	46.1	34.0
CRF-RNN [53]	66.3	48.9	35.4
DeepLab [40]	68.7	46.9	36.8
ACNet [46]	–	–	48.3
MTI-Net [65]	75.3	62.9	49.0
RDFNet [62]	76	62.8	50.1
ESANet [67]	–	–	50.5
ICM [66]	75.4	–	50.7
CANet [72]	76.6	63.8	51.2
ShapeConv [68]	75.8	62.8	51.3
NANet [73]	77.9	–	52.3
SA-Gate [74]	77.9	–	52.4
FFCANet	**78.4**	**65.3**	**53.3**

After determining the network structure, we compared the performance of the best version of FFCANet with similar other work on the publicly available NYUv2 dataset. Table 2 shows the results. Notably, this work achieves remarkable results in several metrics such as Pixel Acc, Mean Acc, and MIoU. This result indicates the advantage of this cross-modal calibration and fusion mechanism in dealing with complex indoor environments with depth images containing noise.

For the pyramid pooling context, using too many combinations of pooling at different resolutions does not boost the network's accuracy. However, it may interfere with the inference process of the subsequent decoder. That may be caused by the small number of feature maps corresponding to a single pooling branch when there are too many pooling branches. Therefore, in the final version of the model, we only use two pooling branches. Their size is 1x1 and 5x3.

4.3 Ablation Experiment

Table 3. Result of ablation experience

Backbone	Encoder	Context	P Acc	M Acc	FW Acc	MIoU
ResNet50	RGB-D	No	75.88	61.43	62.33	49.49
	RGB-D	Yes	76.03	61.79	62.54	49.72
	FFCAM	No	76.34	62.28	63.58	50.74
	FFCAM	**Yes**	**76.96**	**62.71**	**63.9**	**51.19**
ResNet101	RGB-D	No	77.02	62.80	63.82	51.17
	RGB-D	Yes	77.38	62.98	64.52	51.54
	FFCAM	No	77.92	62.85	64.23	52.11
	FFCAM	**Yes**	**78.13**	**63.14**	**64.86**	**52.58**
ResNet152	RGB-D	No	77.52	62.99	63.93	51.54
	RGB-D	Yes	77.58	63.14	64.31	51.87
	FFCAM	No	77.87	64.43	65.27	53.03
	FFCAM	**Yes**	**78.40**	**65.31**	**65.72**	**53.30**

The results of the ablation experiments for the network structure are shown in Table 3, demonstrating the validity of the novel structure of the FFCA Module. A plain RGB-D two-branch segmentation network has been used as the baseline, which removed the FFCA Module between two encoder branches and used a simple element-wise adding instead. The introduction of the global context module also impacts the results, so the global context module is also a variable in the ablation experiments.

This result revealed that the performance of the network is significantly affected by the FFCA Module. When the network uses the FFCA Module as the feature fusion mechanism, there is a significant improvement in the accuracy of the model. Depending on the backbone, this difference can reach approximately 1.4% MIoU. The introduction of global pyramid pooling also contributes

a slight accuracy improvement to the model, with a maximum difference of only approximately 0.3%, which is not as significant as the improvement of the FFCA Module.

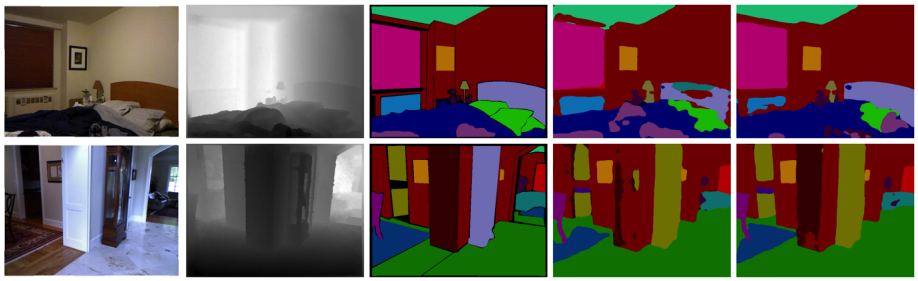

Fig. 7. Visualisation result compared with baseline

The FFCANet with the FFCA Module has obtained better segmentation results than Baseline shown in Fig. 7. As seen from the figure, the segmentation results of FFCANet have fewer category errors, more accurate object edges, and almost no shape breaking. It is due to the FFCA Module's ability to calibrate across modalities and its Squeeze-and-Excitation feature in Channel Attention. This feature allows the module to suppress defects and noise in the depth image very well, acquiring depth information while reducing the interference of harmful parts in the final segmentation result.

5 Conclusion

In this paper, we propose a neural network called FFCANet for accurately executing RGB-D semantic segmentation tasks. We have built a network structure for the semantic segmentation task by modifying the existing ResNet. This module can achieve cross-modal calibration of RGB information with depth information and fuse complementary information. Our experiments show that this ability has made FFCANet get the performance improvement in RGB-D semantic segmentation task.

As the novel structure, the role of FFCA Module is to incorporates two different modalities. This attention module is designed to be plug-and-play, can be combined with any other RGB-D semantic segmentation network have double-branch encoder structure without increasing the burden of calculation. Compared with the baseline in ablation experiment, the model used in this research has obviously improved the accuracy of the semantic segmentation task.

References

1. Thoma, M.: A survey of semantic segmentation. arXiv preprint **arXiv:1602.06541** (2016)
2. 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)
3. Badrinarayanan, V., Kendall, A., Cipolla, R.: Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **39**(12), 2481–2495 (2017)
4. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from RGBD images. In: Fitzgibbon, A.W., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) Computer Vision - ECCV 2012 - 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part V. Lecture Notes in Computer Science, vol. 7576, pp. 746–760. Springer (2012)
5. Quattoni, A., Torralba, A.: Recognizing indoor scenes. In: 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA. pp. 413–420. IEEE Computer Society (2009)
6. Silberman, N., Fergus, R.: Indoor scene segmentation using a structured light sensor. In: IEEE International Conference on Computer Vision Workshops, ICCV 2011 Workshops, Barcelona, Spain, November 6-13, 2011. pp. 601–608. IEEE Computer Society (2011)
7. Webb, J., Ashley, J.: Depth Image Processing, pp. 49–83. Apress, Berkeley, CA (2012)
8. Cai, Z., Han, J., Liu, L., Shao, L.: RGB-D datasets using microsoft kinect or similar sensors: a survey. Multim. Tools Appl. **76**(3), 4313–4355 (2017)
9. Firman, M.: RGBD datasets: Past, present and future. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2016, Las Vegas, NV, USA, June 26 - July 1, 2016. pp. 661–673. IEEE Computer Society (2016)
10. Zhang, Y., Funkhouser, T.A.: Deep depth completion of a single RGB-D image. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018. pp. 175–185. Computer Vision Foundation / IEEE Computer Society (2018)
11. Gupta, S., Girshick, R.B., Arbeláez, P.A., Malik, J.: Learning rich features from RGB-D images for object detection and segmentation. In: Fleet, D.J., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VII. Lecture Notes in Computer Science, vol. 8695, pp. 345–360. Springer (2014)
12. Gupta, S., Arbelaez, P., Malik, J.: Perceptual organization and recognition of indoor scenes from RGB-D images. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, June 23-28, 2013. pp. 564–571. IEEE Computer Society (2013)
13. Hazirbas, C., Ma, L., Domokos, C., Cremers, D.: Fusenet: Incorporating depth into semantic segmentation via fusion-based CNN architecture. In: Lai, S., Lepetit, V., Nishino, K., Sato, Y. (eds.) Computer Vision - ACCV 2016 - 13th Asian Conference on Computer Vision, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part I. Lecture Notes in Computer Science, vol. 10111, pp. 213–228. Springer (2016)
14. Couprie, C., Farabet, C., Najman, L., LeCun, Y.: Indoor semantic segmentation using depth information. In: Bengio, Y., LeCun, Y. (eds.) 1st International Con-

ference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Conference Track Proceedings (2013)
15. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proc. IEEE **86**(11), 2278–2324 (1998)
16. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. In: Bengio, Y., LeCun, Y. (eds.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings (2015)
17. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015. pp. 3431–3440. IEEE Computer Society (2015)
18. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Bartlett, P.L., Pereira, F.C.N., Burges, C.J.C., Bottou, L., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States. pp. 1106–1114 (2012)
19. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S.E., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015. pp. 1–9. IEEE Computer Society (2015)
20. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: Bengio, Y., LeCun, Y. (eds.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings (2015)
21. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016. pp. 770–778. IEEE Computer Society (2016)
22. Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H.: Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint **arXiv:1704.04861** (2017)
23. Zhang, X., Zhou, X., Lin, M., Sun, J.: Shufflenet: An extremely efficient convolutional neural network for mobile devices. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018. pp. 6848–6856. Computer Vision Foundation / IEEE Computer Society (2018)
24. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: Bach, F.R., Blei, D.M. (eds.) Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015. JMLR Workshop and Conference Proceedings, vol. 37, pp. 448–456. JMLR.org (2015)
25. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016. pp. 2818–2826. IEEE Computer Society (2016)
26. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A.: Inception-v4, inception-resnet and the impact of residual connections on learning. In: Singh, S., Markovitch, S. (eds.) Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence,

February 4-9, 2017, San Francisco, California, USA. pp. 4278–4284. AAAI Press (2017)
27. Huang, G., Liu, Z., van der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017. pp. 2261–2269. IEEE Computer Society (2017)
28. Tan, M., Le, Q.V.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: Chaudhuri, K., Salakhutdinov, R. (eds.) Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA. Proceedings of Machine Learning Research, vol. 97, pp. 6105–6114. PMLR (2019)
29. Tan, M., Le, Q.V.: Efficientnetv2: Smaller models and faster training. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event. Proceedings of Machine Learning Research, vol. 139, pp. 10096–10106. PMLR (2021)
30. Cheng, Y., Cai, R., Li, Z., Zhao, X., Huang, K.: Locality-sensitive deconvolution networks with gated fusion for RGB-D indoor semantic segmentation. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017. pp. 1475–1483. IEEE Computer Society (2017)
31. Yu, F., Koltun, V.: Multi-scale context aggregation by dilated convolutions. In: Bengio, Y., LeCun, Y. (eds.) 4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings (2016)
32. Peng, C., Zhang, X., Yu, G., Luo, G., Sun, J.: Large kernel matters - improve semantic segmentation by global convolutional network. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017. pp. 1743–1751. IEEE Computer Society (2017)
33. Chollet, F.: Xception: Deep learning with depthwise separable convolutions. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017. pp. 1800–1807. IEEE Computer Society (2017)
34. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017. pp. 6230–6239. IEEE Computer Society (2017)
35. He, K., Zhang, X., Ren, S., Sun, J.: Spatial pyramid pooling in deep convolutional networks for visual recognition. In: Fleet, D.J., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part III. Lecture Notes in Computer Science, vol. 8691, pp. 346–361. Springer (2014)
36. Yang, M., Yu, K., Zhang, C., Li, Z., Yang, K.: Denseaspp for semantic segmentation in street scenes. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018. pp. 3684–3692. Computer Vision Foundation / IEEE Computer Society (2018)
37. Chen, L., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Semantic image segmentation with deep convolutional nets and fully connected crfs. In: Bengio, Y., LeCun, Y. (eds.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings (2015)
38. Chen, L., 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**(4), 834–848 (2018)

39. Chen, L., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. arXiv preprint **arXiv:1706.05587** (2017)
40. Chen, L., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part VII. Lecture Notes in Computer Science, vol. 11211, pp. 833–851. Springer (2018)
41. Choi, M.J., Lim, J.J., Torralba, A., Willsky, A.S.: Exploiting hierarchical context on a large database of object categories. In: The Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2010, San Francisco, CA, USA, 13-18 June 2010. pp. 129–136. IEEE Computer Society (2010)
42. Mottaghi, R., Chen, X., Liu, X., Cho, N., Lee, S., Fidler, S., Urtasun, R., Yuille, A.L.: The role of context for object detection and semantic segmentation in the wild. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014. pp. 891–898. IEEE Computer Society (2014)
43. Liu, W., Rabinovich, A., Berg, A.C.: Parsenet: Looking wider to see better. arXiv preprint **arXiv:1506.04579** (2015)
44. Hung, W., Tsai, Y., Shen, X., Lin, Z.L., Sunkavalli, K., Lu, X., Yang, M.: Scene parsing with global context embedding. In: IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017. pp. 2650–2658. IEEE Computer Society (2017)
45. Yu, C., Wang, J., Peng, C., Gao, C., Yu, G., Sang, N.: Bisenet: Bilateral segmentation network for real-time semantic segmentation. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XIII. Lecture Notes in Computer Science, vol. 11217, pp. 334–349. Springer (2018)
46. Fu, J., Liu, J., Wang, Y., Li, Y., Bao, Y., Tang, J., Lu, H.: Adaptive context network for scene parsing. In: 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019. pp. 6747–6756. IEEE (2019)
47. Zhang, H., Dana, K.J., Shi, J., Zhang, Z., Wang, X., Tyagi, A., Agrawal, A.: Context encoding for semantic segmentation. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018. pp. 7151–7160. Computer Vision Foundation / IEEE Computer Society (2018)
48. Baltrusaitis, T., Ahuja, C., Morency, L.: Multimodal machine learning: A survey and taxonomy. IEEE Trans. Pattern Anal. Mach. Intell. **41**(2), 423–443 (2019)
49. Fooladgar, F., Kasaei, S.: A survey on indoor RGB-D semantic segmentation: from hand-crafted features to deep convolutional neural networks. Multim. Tools Appl. **79**(7-8), 4499–4524 (2020)
50. Coupé, P., Manjón, J.V., Fonov, V.S., Pruessner, J.C., Robles, M., Collins, D.L.: Patch-based segmentation using expert priors: Application to hippocampus and ventricle segmentation. NeuroImage **54**(2), 940–954 (2011)
51. Wang, M., Liu, X., Gao, Y., Ma, X., Soomro, N.Q.: Superpixel segmentation: A benchmark. Signal Process. Image Commun. **56**, 28–39 (2017)
52. Kaganami, H.G., Zou, B.: Region-based segmentation versus edge detection. In: Pan, J., Chen, Y., Jain, L.C. (eds.) Fifth International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP 2009), Kyoto,

Japan, 12-14 September, 2009, Proceedings. pp. 1217–1221. IEEE Computer Society (2009)
53. Zheng, S., Jayasumana, S., Romera-Paredes, B., Vineet, V., Su, Z., Du, D., Huang, C., Torr, P.H.S.: Conditional random fields as recurrent neural networks. In: 2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015. pp. 1529–1537. IEEE Computer Society (2015)
54. Banica, D., Sminchisescu, C.: Second-order constrained parametric proposals and sequential search-based structured prediction for semantic segmentation in RGB-D images. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015. pp. 3517–3526. IEEE Computer Society (2015)
55. Bo, L., Ren, X., Fox, D.: Kernel descriptors for visual recognition. In: Lafferty, J.D., Williams, C.K.I., Shawe-Taylor, J., Zemel, R.S., Culotta, A. (eds.) Advances in Neural Information Processing Systems 23: 24th Annual Conference on Neural Information Processing Systems 2010. Proceedings of a meeting held 6-9 December 2010, Vancouver, British Columbia, Canada. pp. 244–252. Curran Associates, Inc. (2010)
56. Hermans, A., Floros, G., Leibe, B.: Dense 3d semantic mapping of indoor scenes from RGB-D images. In: 2014 IEEE International Conference on Robotics and Automation, ICRA 2014, Hong Kong, China, May 31 - June 7, 2014. pp. 2631–2638. IEEE (2014)
57. Lerma, C.D.C., Kosecká, J.: Semantic parsing for priming object detection in indoors RGB-D scenes. Int. J. Robotics Res. **34**(4-5), 582–597 (2015)
58. Ren, X., Bo, L., Fox, D.: RGB-(D) scene labeling: Features and algorithms. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, June 16-21, 2012. pp. 2759–2766. IEEE Computer Society (2012)
59. Müller, A.C., Behnke, S.: Learning depth-sensitive conditional random fields for semantic segmentation of RGB-D images. In: 2014 IEEE International Conference on Robotics and Automation, ICRA 2014, Hong Kong, China, May 31 - June 7, 2014. pp. 6232–6237. IEEE (2014)
60. Wang, S., Lokhande, V.S., Singh, M., Körding, K.P., Yarkony, J.: End-to-end training of CNN-CRF via differentiable dual-decomposition. arXiv preprint **arXiv:1912.02937** (2019)
61. McCormac, J., Handa, A., Leutenegger, S., Davison, A.J.: Scenenet RGB-D: can 5m synthetic images beat generic imagenet pre-training on indoor segmentation? In: IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017. pp. 2697–2706. IEEE Computer Society (2017)
62. Lee, S., Park, S., Hong, K.: Rdfnet: RGB-D multi-level residual feature fusion for indoor semantic segmentation. In: IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017. pp. 4990–4999. IEEE Computer Society (2017)
63. Lin, G., Milan, A., Shen, C., Reid, I.D.: Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017. pp. 5168–5177. IEEE Computer Society (2017)
64. Li, Z., Gan, Y., Liang, X., Yu, Y., Cheng, H., Lin, L.: LSTM-CF: unifying context modeling and fusion with lstms for RGB-D scene labeling. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II. Lecture Notes in Computer Science, vol. 9906, pp. 541–557. Springer (2016)

65. Vandenhende, S., Georgoulis, S., Gool, L.V.: Mti-net: Multi-scale task interaction networks for multi-task learning. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J. (eds.) Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part IV. Lecture Notes in Computer Science, vol. 12349, pp. 527–543. Springer (2020)
66. Shi, H., Li, H., Wu, Q., Song, Z.: Scene parsing via integrated classification model and variance-based regularization. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019. pp. 5307–5316. Computer Vision Foundation / IEEE (2019)
67. Seichter, D., Köhler, M., Lewandowski, B., Wengefeld, T., Gross, H.: Efficient RGB-D semantic segmentation for indoor scene analysis. In: IEEE International Conference on Robotics and Automation, ICRA 2021, Xi'an, China, May 30 - June 5, 2021. pp. 13525–13531. IEEE (2021)
68. Cao, J., Leng, H., Lischinski, D., Cohen-Or, D., Tu, C., Li, Y.: Shapeconv: Shape-aware convolutional layer for indoor RGB-D semantic segmentation. In: 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021. pp. 7068–7077. IEEE (2021)
69. Woo, S., Park, J., Lee, J., Kweon, I.S.: CBAM: convolutional block attention module. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part VII. Lecture Notes in Computer Science, vol. 11211, pp. 3–19. Springer (2018)
70. Li, X., Wang, W., Hu, X., Yang, J.: Selective kernel networks. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019. pp. 510–519. Computer Vision Foundation / IEEE (2019)
71. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018. pp. 7132–7141. Computer Vision Foundation / IEEE Computer Society (2018)
72. Zhou, H., Qi, L., Huang, H., Yang, X., Wan, Z., Wen, X.: Canet: Co-attention network for RGB-D semantic segmentation. Pattern Recognit. **124**, 108468 (2022)
73. Zhang, G., Xue, J., Xie, P., Yang, S., Wang, G.: Non-local aggregation for RGB-D semantic segmentation. IEEE Signal Process. Lett. **28**, 658–662 (2021)
74. Chen, X., Lin, K., Wang, J., Wu, W., Qian, C., Li, H., Zeng, G.: Bi-directional cross-modality feature propagation with separation-and-aggregation gate for RGB-D semantic segmentation. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J. (eds.) Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XI. Lecture Notes in Computer Science, vol. 12356, pp. 561–577. Springer (2020)