

Information Fusion

BCINet: Bilateral Cross-Modal Interaction Network for Indoor Scene Understanding in RGB-D Images

--Manuscript Draft--

Manuscript Number:	INFFUS-D-22-00538
Article Type:	Research paper
Keywords:	Scene understanding, RGB-D, bilateral cross-modal interaction, hybrid pyramid dilated convolution, deep learning.
Abstract:	<p>Depth cue has proven to be useful information in the indoor scene understanding of RGB-D images for providing a geometric counterpart to RGB representation. However, because of the differences between RGB-D image pairs, utilizing cross-modal data effectively is a key issue. Most methods exclusively leverage depth data to unilaterally complement RGB data for better feature representation; they invariably ignore the fact that RGB and depth data can bilaterally complement each other. Herein, a novel RGB-D scene-understanding network called BCINet is presented, in which RGB and depth data bilaterally complement each other via a proposed bilateral cross-modal interaction module (BCIM). The BCIM helps to capture cross-modal complementary cues by crossly fusing enhanced features from one modality to the counterpart modality through a feature enhanced module. Meanwhile, exploiting the long-range dependencies of RGB-D features is also significant for accurate scene understanding. Specifically, we design a hybrid pyramid dilated convolution module to enlarge the receptive fields along both the vertical and horizontal spatial directions to adaptively capture diverse contexts with different shapes. Additionally, we propose a context-guided module to aggregate these diverse higher-level contexts with lower-level features in the encoder to guide the information flow for progressively refining the segmentation map. Experimental results on two indoor scene datasets demonstrate the superiority and effectiveness of the proposed BCINet over several state-of-the-art approaches.</p>

Highlights

- 1) Bilateral cross-modal interaction is achieved and hybrid contexts are exploited to improve segmentation accuracy.
- 2) A proposed FEM is embedded to highlight unimodally useful features.
- 3) An HPDC is designed to effectively collect diverse contextual information for improving the capability of BCINet.
- 4) A CGM is designed to transmit contextual information.

BCINet: Bilateral Cross-Modal Interaction Network for Indoor Scene Understanding in RGB-D Images

Wujie Zhou^{a,b*}, Yuchun Yue^a, Meixin Fang^b, Xiaohong Qian^a, Lu Yu^b

^a *School of Information & Electronic Engineering, Zhejiang University of Science & Technology, Hangzhou 310023, China*

^b *College of Information Science & Electronic Engineering, Zhejiang University, Hangzhou 310027, China*

Abstract

Depth cue has proven to be useful information in the indoor scene understanding of RGB-D images for providing a geometric counterpart to RGB representation. However, because of the differences between RGB-D image pairs, utilizing cross-modal data effectively is a key issue. Most methods exclusively leverage depth data to unilaterally complement RGB data for better feature representation; they invariably ignore the fact that RGB and depth data can bilaterally complement each other. Herein, a novel RGB-D scene-understanding network called BCINet is presented, in which RGB and depth data bilaterally complement each other via a proposed bilateral cross-modal interaction module (BCIM). The BCIM helps to capture cross-modal complementary cues by crossly fusing enhanced features from one modality to the counterpart modality through a feature enhanced module. Meanwhile, exploiting the long-range dependencies of RGB-D features is also significant for accurate scene understanding. Specifically, we design a hybrid pyramid dilated convolution module to enlarge the receptive fields along both the vertical and horizontal spatial directions to adaptively capture diverse contexts with different shapes. Additionally, we propose a context-guided module to aggregate these diverse higher-level contexts with lower-level features in the encoder to guide the information flow for progressively refining the segmentation map. Experimental results on two indoor scene datasets demonstrate the superiority and effectiveness of the proposed BCINet over several state-of-the-art approaches.

Keywords: Scene understanding, RGB-D, bilateral cross-modal interaction, hybrid pyramid dilated convolution, deep learning.

1. Introduction

Scene understanding aims to assign a category label to each pixel in an image; thus, it is a fundamental and essential machine vision task for many applications such as video surveillance,

*Corresponding author.

E-mail addresses: wujiezhou@163.com

Postal address: 310023

Tel.: +86-571-85070303.

medical image analysis, robot perception, and, autonomous driving [1–3]. RGB data contain rich appearance information and texture details, whereas depth data provide useful geometric cues for reliable segmentation of objects with ambiguous appearances [4]. With the development of range sensors (such as LiDAR and Kinect), depth data are applicable along with RGB images. By exploiting the two complementary modalities with a convolutional neural network (CNN) the scene-understanding performance can be improved.

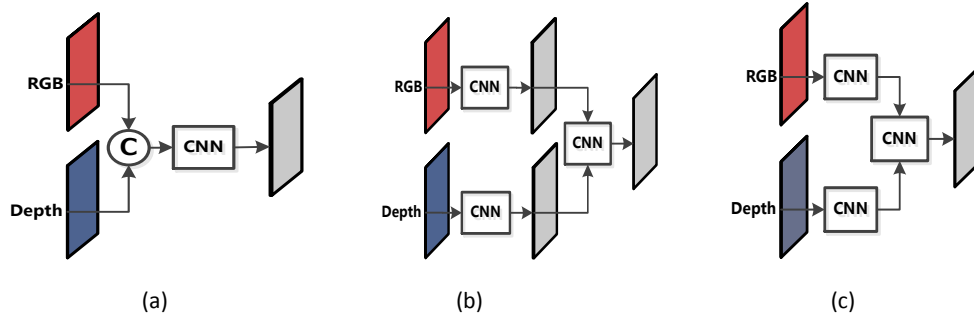


Fig. 1. Typical architectures to explore the relationship between an RGB image and depth map for scene understanding. (a) Concatenating RGB and depth channels, (b) fusing score maps computed from each modality, (c) sending RGB images and depth maps to two streams, then performing weighted fusion on the feature maps of both streams.

Concomitant with the success of vision recognition using CNNs, many scene-understanding approaches have adopted CNNs by extending them to pixel-wise classification. To solve the main problem of indoor scene understanding in RGB-D images of how to effectively extract and merge depth features along with RGB features, various methods have been presented that exploit the ability of CNNs for fusing depth data. The methods include concatenating the input RGB and depth streams, fusing score maps calculated from each modality, and sending RGB images and depth data to two streams, then performing weighted fusion on the feature maps of both streams, as shown in Figs. 1(a)–(c), respectively. Although these approaches achieve meaningful results, the correlation and interaction between RGB and depth data are only implicitly scrutinized. Furthermore, existing studies typically unilaterally assert that depth data are generally noisy, and may not lead to satisfactory results for scene understanding.

In contrast to the above assertion, we argue that the RGB image may also contain noisy features caused by the similar appearance of different neighboring objects, a possibility that is completely ignored by most studies. Neither RGB data nor depth data are completely flawless to their counterpart modality; they should be complemented with each other instead of only supplying depth map to RGB image data unilaterally. How to exploit the beneficial properties of the two types of data mutually and bilaterally is still rarely explored in existing studies.

Based on the above analysis, we propose a bilateral cross-modal interaction module (BCIM) to achieve bilateral interaction and to better exploit the relation between RGB and depth data. The key idea is to use a proposed feature enhanced module (FEM) to extract significant unimodal feature

information by using attention mechanisms from both spatial-wise and channel-wise dimensions. Then, these unimodal useful features are crossly merged with their counterpart modal features. By merging discriminative features from the opposite modality, the BCIM facilitates bilateral cross-modal complementary information interaction without deteriorating modality-specific feature extraction, which promotes higher segmentation accuracy.

Simultaneously, context plays an important role in scene understanding, i.e., low-level contextual information could provide fine details for delineating the objects' boundary, whereas high-level contextual information describes the relationship between different parts and thus is extremely helpful for discovering the specific locations of objects. Diverse contexts contain more adaptive information of objects with irregular shapes in the complex indoor scenes. However, the issue of how to learn diverse contextual representation is still understudied in existing studies.

To explore more contexts for achieving high performance, atrous convolution [5] has proven highly effective in capturing contexts for classification tasks. However, a common limitation for regular dilated convolutions applied in atrous spatial pyramid pooling (ASPP) is that they all probe the input feature maps within square windows. This limits their flexibility in capturing anisotropic contexts that exist extensively in realistic scenarios. For example, the target objects may have a long-range strip structure in some cases, whether along the horizontal or vertical direction (e.g., the pillar and shelf in the first row of Fig. 2, the cabinet in the second row, and the pictures in the third row).

Merely using large square dilated convolution windows cannot adequately address this problem since it would incorporate contaminating cues from irrelevant regions [6]. Therefore, we propose a hybrid pyramid dilated convolution (HPDC) in this study, where we integrate rectangular dilated convolution into the common atrous convolution framework to collect more specific and diverse contexts apart from general contexts. The proposed HPDC deploys a long kernel shape along one spatial dimension with variable dilated rates but a narrow kernel shape along the other spatial dimension, which enables capturing of the long-range relationships of isolated regions, as shown in the third column of Fig. 2. This is essentially different from the traditional atrous convolution that only collects contextual information from a fixed square region, shown in the second column of Fig. 2. Furthermore, to successively propagate the diverse contexts collected by HPDC from higher to lower levels and effectively leverage the information of the encoder, we designed a context-guided module (CGM) in the decoder to harvest the flexible contexts of different levels.

The main contributions of this work are as follows:

- 1) A novel two-stream RGB-D scene-understanding network called BCINet is introduced, in which bilateral cross-modal interaction is achieved and hybrid contexts are exploited to improve segmentation accuracy.
- 2) A BCIM is introduced to bilaterally capture the cross-modal complementary cues between the unimodal RGB images and depth data, where a proposed FEM is embedded to highlight unimodally useful features.
- 3) An HPDC is designed to effectively collect diverse contextual information for improving the

capability of BCINet to incorporate more anisotropic and characteristic features. A CGM is also designed to transmit contextual information collected by the HPDC from high levels to lower-level features for more diverse contextual exploration.

4) The proposed BCINet achieves state-of-the-art (SOTA) performances on two indoor scene datasets.

2. Related Work

2.1 RGB-D Scene Understanding

Recently, scene understanding based on fully convolutional networks [7] has achieved high performance. The introduction of the Microsoft Kinect sensor has significantly simplified the acquisition of depth data, and numerous segmentation methods based on RGB and depth modalities and their fusions have been developed. For instance, the methods in [8, 9] adopted an early fusion scheme for RGB and depth data for segmentation. Long *et al.* [7] used late fusion by adding the prediction results from the RGB and depth modalities. Peng *et al.* [10] directly concatenated RGB and depth data into a four-stream input to a multistage model. Nevertheless, the fusions in the aforementioned methods are mainly conducted by simple addition or concatenation; they fail to reflect the relations between RGB and depth data accurately.

To enhance fusions between the two types of data, subsequent methods more accurately exploit data relations. Han *et al.* [11] used a dual-channel late fusion network to integrate deep feature maps from RGB and depth branches by appending a shared layer. Cheng *et al.* [12] adopted a gated fusion network to learn the varying contributions of each modality to determine categories in different scenes. Xiong *et al.* [13] leveraged a variational context-deformable model for structured learning of adaptive receptive fields. Ma *et al.* [14] performed scene understanding by applying deep learning to RGB-D data in a multi-view context. Liu *et al.* [15] performed joint modeling by learning a weighted sum model to fuse the results from two-stream deep CNNs using RGB and depth encoding. He *et al.* [16] provided a superpixel-based network for scene understanding. First, region correspondences were computed from the optical flow and image boundary superpixels, and then a novel spatiotemporal pooling layer aggregated the corresponding information. Fayyaz *et al.* [17] included spatial and temporal features for scene understanding of street scenes by combining a long short-term memory (LSTM) network and a CNN. Liu *et al.* [18] introduced a fully connected deep continuous conditional random field (CRF) module for continuous and discrete labeling. Lin *et al.* [19] considered depth information to split an image into different layers with common characteristics regarding scenes/objects; a context-aware receptive field improved control of relevant contexts in the learned deep features. Wang *et al.* [20] proposed a depth-aware network comprising depth-aware average pooling and depth-aware convolutions. The scene geometry was seamlessly incorporated into the CNN by leveraging the depth similarity between pixels during information propagation. Yuan *et al.* [21] achieved indoor scene segmentation by incorporating RGB-D data and using a

squeeze-and-excitation residual framework as an encoder to extract feature maps from RGB-D data simultaneously. Then, feature fusion was performed during decoding. Lin *et al.* [22] introduced a switchable context model to facilitate indoor scene understanding in RGB-D images. Their model analyzed the information in image regions to identify representative characteristics. Depth data allowed the identification of objects in these regions. Zhou *et al.* [23] performed indoor scene understanding using their three-branch self-attention framework that comprised two asymmetric input branches (i.e., asymmetric encoders) and a cross-modality distillation branch with a self-attention block. Wang *et al.* [24] used a feature transformation architecture to bridge deconvolutional and convolutional modules. They correlated the RGB and depth modalities by extracting common features; the modalities were characterized by specific features. Zhang *et al.* [25] proposed a multi-modality non-local aggregation network, which can effectively fuse non-local RGB-D feature maps along different dimensions. Li *et al.* [26] introduced the LSTM context fusion model to capture the context in RGB and depth streams by stacking an LSTM layer and several convolutional layers, respectively. Jiang *et al.* [27] presented a residual encoder–decoder framework for indoor scene understanding in RGB-D images. Qi *et al.* [28] introduced a graph neural module that builds a *KNN* graph on top of a point cloud. Lee *et al.* [29] presented a multimodality feature block and multilevel feature refinement block to fuse multilevel features through a top-down path. Similarly, Valada *et al.* [30] used self-supervised model adaptation fusion to combine modality-specific streams and fuse multilevel features through a top-down path.

Some methods consider depth information as an additional supervised signal to recalibrate RGB data in multitask learning. For instance, Zhang *et al.* [31] presented a pattern affinity propagation architecture to boost and regularize complementary tasks. Xu *et al.* [32] introduced a multimodality distillation module to pass valid information from depth data to RGB features. Chen *et al.* [33] presented an SA-Gate unit to guarantee cross-modality feature maps aggregation via an attention mechanism.

Although the abovementioned methods use complicated architectures and have notably advanced scene understanding, the novelty of our BCINet is that our model explicitly formulates bilateral interdependencies in a cross-modality framework without deteriorating modality-specific feature extraction, instead of unilaterally sending depth data to complement the RGB image or designing a complicated architecture to fuse the two types of data for a unified representation.

2.2 Attention Mechanisms

The attention mechanism resembling that of the human visual system has evolved along with some variants to perform many computer vision tasks. For instance, Chen *et al.* [34] used a cross-modality cross-level attention mechanism to aggregate useful cues from RGB and depth data while discarding redundant information. Zhang *et al.* [35] and Liu *et al.* [36] combined attention mechanisms with U-shaped models to guide feature integration. Yuan *et al.* and Fu *et al.* [37, 38] used self-attention module to obtain long-range contexts, obtaining a pixel-wise similarity results

based on pairs of semantics. Jia *et al.* [39] aggregated global contexts by learning a position-sensitive spatial attention mask to guide information flow. Wang *et al.* [40] introduced a non-local network for learning self-attention in 2D/3D vision modeling. Huang *et al.* [41] replaced a densely connected attention stream with a crisscross stream to improve network efficiency. Cao *et al.* [42] unified the non-local architecture with SENet [43] to learn a global context using a lightweight architecture. Zhou *et al.* [44] presented a convolutional gated recurrent unit to extract valuable information iteratively from high levels while jointly using channel attention and spatial attention to distill discriminant cues from low levels. Hu *et al.* [45] introduced the attention complementary framework that gathers deep features from RGB-D data. Then, a channel-attention-based model extracts weighted features from these streams. Deng *et al.* [46] presented a residual fusion module with an attention mechanism. Zhang *et al.* [47] proposed an adaptive fusion module that explores the contributions of fused deep features at different levels to impede semantic dilution at high levels by applying feature-wise attention to learn weights exclusively. Zhou *et al.* [48] proposed a co-attention fusion method, which used an RGB feature query and a set of depth feature key-value pairs to transform the depth feature into the RGB feature space and then fuse with the local RGB features.

In BCINet, we designed an FEM to highlight discriminative unimodal features through attention mechanisms. The FEM contains four parallel branches from both channel-wise and spatial-wise dimensions, each with a different focus.

2.3 Context Learning

Context plays a very important role in achieving high performance for various computer vision tasks. Many models are devoted to exploiting more discriminative contexts to improve their feature representation ability. Early techniques for modeling contextual relations for scene understanding involved CRFs [49, 50]. However, CRFs are mostly modeled in the discrete label, and thus are now less successful for producing SOTA results of scene parsing, even though they have been integrated into CNNs (strip pooling). Chen *et al.* [5] introduced an ASPP to capture multi-scale contexts for scene understanding. Zhao *et al.* [51] employed a pyramid pooling model to aggregate global and local contexts. Mei *et al.* [52] adopted two parallel context exploration (CE) blocks to perceive rich contextual information over a wide range of scales. Zhang *et al.* [53] proposed a multi-contextual module to distill features with different receptive fields and contextual characteristics. However, the ability of these methods to exploit contexts is limited because only the regular square kernel shapes are applied.

Our method differs from the above studies in that it explores diverse contextual information in an HPDC that integrates rectangular dilated convolutions into the common atrous convolution framework. The proposed HPDC module helps to enhance the discriminability of the framework via hybrid multi-scale contexts.

3. Proposed BCINet

3.1 Overview of the Architecture

Fig. 3 shows the framework of the proposed encoder–decoder BCINet, which takes RGB images and depth data as inputs and outputs the segmentation map in an end-to-end manner. The proposed BCINet is designed as an encoder–decoder architecture because that type of network is able to effectively obtain high-level semantic information as well as low-level detail information. For the encoder, we chose ResNet-50 [54] as the feature extractor for both the RGB and depth branches. After the RGB and depth data are fed into the two ResNet-50 branches, five levels of side-output feature maps of each modality are obtained, F_i^R for RGB images and F_i^D for depth data ($i=1, 2, 3, 4, 5$). At the first four stages of the RGB and depth branches, bilateral cross-modal feature interaction is achieved via the proposed BCIM, in which the proposed FEM is employed to exploit unimodally specific features, and these features are subsequently fused with the counterpart modal features. We fuse the features of each stage of both modalities by element-wise addition, denoted as $\{M_i \mid i=1, 2, 3, 4, 5\}$, which are later fed to the CGM by skip connection. For the decoder part, the HPDC and CGM are employed in succession at each stage, where the HPDC module is first leveraged to explore diverse higher-level contextual information. The results and M_i are then jointly fed to the CGM to successively aggregate these diverse contexts with the lower-level features for better information flow. Finally, we use a convolution layer with a 3×3 kernel and an ensuing upsample operation on the feature maps of the last decoder to predict the final segmentation map.

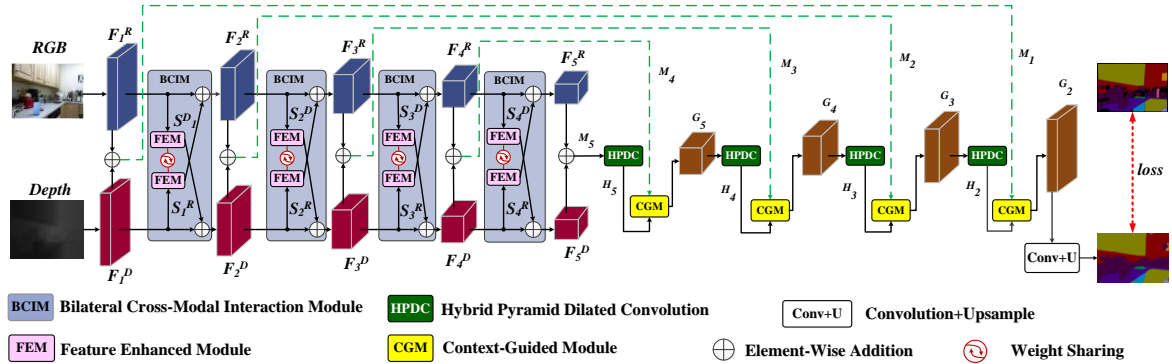


Fig. 3. Overall architecture of the proposed BCINet for RGB-D indoor scene understanding.

3.2 Bilateral Cross-Modal Interaction Module (BCIM)

RGB-D scene understanding is a challenging computer vision task because there are large modality gaps between the RGB and depth data. Most existing methods unilaterally complement depth data to image data or simply fuse the unimodal RGB and depth features by concatenation and element-wise addition for capturing cross-modal information. Although these models may work well in some cases, they may produce some undesirable results when the relationships between RGB images and depth data are complicated. To solve this problem, a BCIM is presented, where we design an FEM to explore

one-side unimodal enhanced features and then crossly feed them to the other-side unimodal features to achieve bilateral cross-modal interactions between both modalities.

The detailed framework of BCIM can be seen in Fig. 3, where each BCIM is shown as a light-blue block with two FEM submodules. Given the unimodal feature maps $F_i^R \in \mathbb{R}^{C \times H \times W}$ and $F_i^D \in \mathbb{R}^{C \times H \times W}$ at the i -th level ($i = 1, 2, 3, 4$), we utilize the FEM to extract unimodal distinguishing features S_i^R and S_i^D , respectively, for RGB and depth images. Subsequently, these discriminative features are crossly merged into the counterpart modality by element-wise addition to obtain cross-modal features, which constitute the input to the next stage:

$$F_{i+1}^R = F_i^R + S_i^D, \quad i = 1, 2, 3, 4 \quad (1)$$

$$F_{i+1}^D = F_i^D + S_i^R, \quad i = 1, 2, 3, 4 \quad (2)$$

Both the RGB branch and depth branch adopt an FEM; meanwhile, the FEMs at the same stage applied in two modalities share the same weights, as shown in Fig. 3.

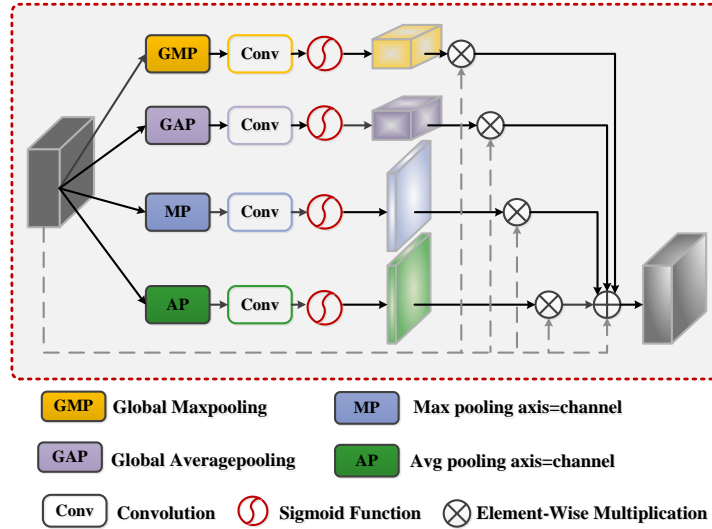


Fig. 4. Detailed structure of the proposed FEM.

The details of the FEM are presented in Fig. 4. Suppose that the input to the FEM is F_i^R ; we design four parallel branches to highlight the distinguishing parts of the input features from both spatial-wise and channel-wise dimensions. The first two branches respectively perform global max pooling and global average pooling operations on the input features along the spatial direction. The results are subsequently sent to a 3×3 convolutional layer and a sigmoid function to obtain a channel-wise attention map. The channel-wise attention features are then pixel-wise multiplied with the input features to generate the channel weighted features. Similarly, the last two branches respectively perform max and average pooling operations on the input features along the channel direction, and are subsequently fed to a 3×3 convolutional layer and a sigmoid function to obtain

spatial-wise attention features. The generated spatial attention maps are then pixel-wise multiplied with the input feature maps to obtain the spatially weighted feature maps. The above process can be formulated as follows:

$$\begin{cases} B_i^1 = \sigma(\text{conv}(\text{GMP}(F_i^R))) \times F_i^R \\ B_i^2 = \sigma(\text{conv}(\text{GAP}(F_i^R))) \times F_i^R \\ B_i^3 = \sigma(\text{conv}(\text{MP}(F_i^R))) \times F_i^R \\ B_i^4 = \sigma(\text{conv}(\text{AP}(F_i^R))) \times F_i^R \end{cases}, i = 1, 2, 3, 4 \quad (3)$$

where *GMP*, *GAP*, *MP*, and *AP* respectively represent global max pooling, global average pooling, max pooling, and average pooling. Conv denotes a 3×3 convolutional layer, σ denotes the sigmoid function, and \times denotes pixel-wise multiplication. B_i^j is the result of the j -th branch ($j = 1, 2, 3, 4$) at the i -th level. The four enhanced features processed by each parallel branch are ultimately fused with the original input features by element-wise addition for better information incorporation:

$$S_i^R = B_i^1 + B_i^2 + B_i^3 + B_i^4 + F_i^R, \quad i = 1, 2, 3, 4 \quad (4)$$

Our FEM leverages four parallel branches to highlight the specific features of unimodality with each having a different focus, effectively enhancing the unimodal feature representation. With the proposed FEM embedded in the BCIM, bilateral interactions between unimodal RGB and depth feature maps are explored to capture effective cross-modal complementary cues, which explicitly formulate the correlations between the two types of data.

3.3 Hybrid Pyramid Dilated Convolution (HPDC) and Context-Guided Module (CGM)

The objects in indoor datasets may exhibit irregular scale changes. This necessitates features with diverse scale to cover the large-scale range for inferring various segmentation objects. Chen *et al.* [5] introduced an ASPP for generating multi-scale feature maps. More recently, Yang *et al.* [55] developed DenseASPP to densely cover the scale range by connecting atrous convolution layers in a dense manner [56]. Wang *et al.* [57] introduced a parallel ASPP module by imposing additional dilated convolutional layers into each dilated layer of the ASPP. Although these methods can obtain multi-scale feature maps that cover a large receptive field, their ability to exploit contexts is limited because only square kernel shapes are applied and they may unavoidably incorporate many irrelevant regions when processing objects with irregular shapes. Therefore, we propose an HPDC module to adaptively collect more diverse and anisotropic contextual information. Additionally, to connect the diverse contextual information with lower-level features in the encoder, we design a context-guided module (CGM) to progressively refine the segmentation map.

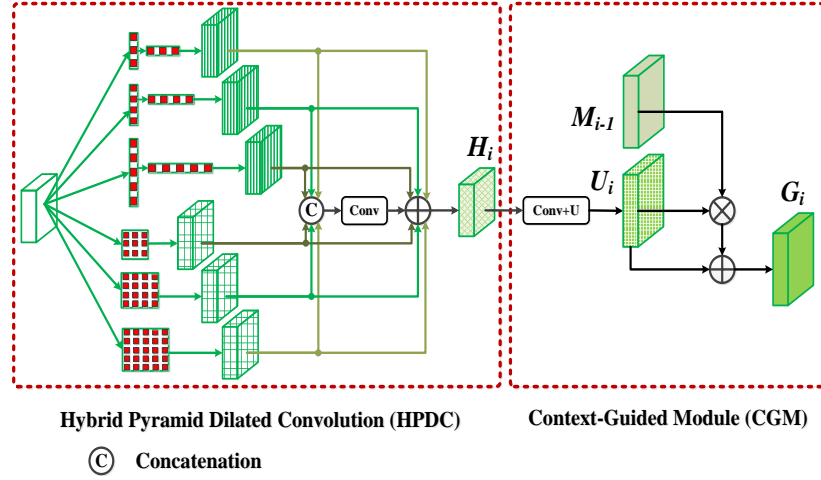


Fig. 5. Detailed structure of the proposed HPDC and CGM.

As shown in the first half of Fig. 5, the HPDC consists of six parallel dilated convolutional branches with specific settings. Each of the first three branches adopts a kernel size of $1 \times N$ and a following $N \times 1$ with dilation rates of r , $2r$, and $3r$, respectively, which are tailored to collect anisotropic contexts with irregular shapes. Concurrently, the other three parallel dilated convolutional branches with square kernel shapes are also applied to preserve the capability for common feature extractions. They adopt a unified kernel size of 3×3 and have dilated rates of k , $2k$, and $3k$, respectively. Suppose that the input of HPDC is X ; the process can be formulated as follows:

$$\begin{cases}
 X_1 = \text{Conv}_{3 \times 1}^r(\text{Conv}_{1 \times 3}^r(X)) \\
 X_2 = \text{Conv}_{5 \times 1}^{2r}(\text{Conv}_{1 \times 5}^{2r}(X)) \\
 X_3 = \text{Conv}_{7 \times 1}^{3r}(\text{Conv}_{1 \times 7}^{3r}(X)) \\
 X_4 = \text{Conv}_{3 \times 3}^k(X) \\
 X_5 = \text{Conv}_{3 \times 3}^{2k}(X) \\
 X_6 = \text{Conv}_{3 \times 3}^{3k}(X) \\
 X_c = \text{Cat}(X_1, X_2, X_3, X_4, X_5, X_6)
 \end{cases} \quad (5)$$

where Conv represents the convolution with different kernel size, as illustrated by the subscripts in Eq. (5). Cat is the concatenation operation. The parameter N is set as 3, 5, and 7, respectively, for the first three branches; r is set as 2, 4, 4, and 2 and k is 6, 4, 2, and 1, respectively, for different levels of the decoder. Then, we apply a 1×1 convolution to squeeze the concatenation results to the same number as the input and they are further fused with the contextual features of the six parallel branches:

$$\begin{cases}
 T_i = \text{ReLU}(\text{BN}(\text{Conv}_{1 \times 1}(X_c))) & i=2, 3, 4, 5 \\
 H_i = T_i + X_1 + X_2 + X_3 + X_4 + X_5 + X_6
 \end{cases} \quad (6)$$

where BN denotes the batch normalization and ReLU is the activation function. H_i are the features output from the HPDC module, which are subsequently fed to the CGM, as displayed in the second

half of Fig. 5. Mathematically, given H_i and M_{i-1} , the process of CGM is calculated by the following:

$$U_i = \text{Up}(\text{ReLU}(\text{BN}(\text{Conv}_{1 \times 1}(H_i)))) , \quad i=2, 3, 4, 5 \quad (7)$$

$$G_i = U_i \times M_{i-1} + U_i , \quad i=2, 3, 4, 5 \quad (8)$$

where Up represents the upsample operation and G_i is the result output from CGM. With the proposed HPDC and CGM collaborating, more diverse and anisotropic contexts are exploited apart from the common contexts. Meanwhile, these higher-level contexts are fused with lower-level features in the encoder, achieving more robust and adaptive feature representation for better boundary delineation. The difference between our HPDC and the common ASPP is that we develop dilated rectangular convolutions with specific sizes to capture more anisotropic and diverse contexts, especially for those objects with irregular shapes. Simultaneously, the dilated rectangular convolutions we leverage can help to better cover the long dependencies along both the horizontal and vertical dimensions through our dilated settings, which is also different from the ordinary rectangular convolutions in some existing methods. The effectiveness of the design of our HPDC is validated in the ensuing ablation studies presented below.

4. Experimental Results

A. Datasets and Evaluation

To measure the prediction performance of the proposed BCINet, we conducted extensive experiments on two publicly used RGB-D scene-understanding datasets: NYUv2 [58] and SUN RGB-D [59]. The NYUv2 and SUN RGB-D (SUN for short) datasets contain 1,449 and 10,335 RGB-D images, respectively, and are divided into training and testing sets. We measured the proposed BCINet with existing SOTA approaches according to three measures: mean accuracy (mAcc.), pixel accuracy (PixAcc.), and mean intersection over union (mIoU).

B. Implementation Details

We implemented our proposed BCINet using the publicly available PyTorch library and executed it on an NVIDIA TITAN V GPU with 12 GB of memory in total. During training, we applied synchronous transforms to the inputs, including the RGB image, depth map, and corresponding annotations, for data augmentation. The operations included random flipping, random scaling, random cropping, random brightness adjustment, contrast adjustment, and saturation adjustment. We used ResNet-50s pre-trained on ImageNet [60] as our backbone in the encoder. The training and testing images were resized to 480×640 resolutions. The SGD was applied as our optimizer, with weight decay 0.0005 and momentum 0.9. We used a batch size of 4/2 and trained the proposed BCINet for 150 and 100 epochs on the NYUv2 and SUN datasets, respectively. The learning rate was adjusted by the poly strategy [60] with a power of 0.9 and a basic learning rate of 0.005. We employed the cross-entropy loss function between the final predicted segmentation map and ground truth for end-to-end BCINet training. For the SUN dataset, when comparing with the SOTA models, we adopted flipping and multi-scale inference strategies for test-time augmentation and added the background as an additional weighted

class during the training process to boost the performance.

C. Comparison with SOTA Methods

1) *Quantitative Comparison on NYUv2 Dataset:* We compared the experimental results of the proposed BCINet with those of the SOTA approaches in [7, 14–23, 45, 48]. The experiments were conducted on the NYUv2 dataset. As illustrated in Table 1, we achieved superior results on the three evaluated measures. We attribute the better performance to the proposed BCIM, HPDC, and CGM, which respectively play a very important role in the accuracy of the entire framework. On the most important metric, mIoU, we achieved 52.95%—an approximately 0.55% improvement over the recent SOTA method by Chen *et al.* [33].

Table 1 Performance of different methods on the NYUv2 dataset.

Models	mAcc	PixAcc	mIoU
Long <i>et al.</i> [7]	46.1	65.4	34.0
Fayyaz <i>et al.</i> [17]	42.6	62.1	30.9
Liu <i>et al.</i> [18]	39.0	63.1	29.5
Ma <i>et al.</i> [14]	51.78	70.66	40.07
Liu <i>et al.</i> [15]	51.7	70.3	41.2
He <i>et al.</i> [16]	53.8	70.1	40.1
Wang <i>et al.</i> [20]	56.3	-	43.9
Lin <i>et al.</i> [19]	-	-	47.7
Yuan <i>et al.</i> [21]	59.27	74.42	46.79
Zhou <i>et al.</i> [23]	59.6	73.50	46.1
Hu <i>et al.</i> [45]	-	-	48.3
Xiong <i>et al.</i> [13]	63.5	-	50.7
Lin <i>et al.</i> [22]	-	-	50.7
Chen <i>et al.</i> [33]	-	77.9	52.4
Zhou <i>et al.</i> [48]	63.8	76.6	51.2
Zhang <i>et al.</i> [25]	-	77.9	52.3
Ours	66.78	77.17	52.95

2) *Quantitative Comparison on SUN Dataset:* We performed additional experiments on the SUN dataset to examine the effectiveness of the proposed BCINet. The SUN dataset contains images from several different datasets. Compared with the NYUv2 dataset, the SUN dataset has more complex scenes and conditions, which we deemed suitable to evaluate the generality of BCINet. From this dataset, we selected 5,285 RGB-D images for training and used the remaining RGB-D images for testing. We compared BCINet with the 13 SOTA scene-understanding methods in [4, 16, 19–20, 22, 25, 27–29, 33, 48] on the SUN dataset. As shown in Table 2, BCINet outperformed the SOTA methods.

Table 2 Performance of different models on the SUN dataset.

Models	mAcc	PixAcc	mIoU
Hazirbas <i>et al.</i> [4]	48.3	76.27	37.29
He <i>et al.</i> [16]	41.2	65.5	32.9
Wang <i>et al.</i> [20]	53.5	-	42.0
Jiang <i>et al.</i> [27]	60.3	81.3	47.8
Lin <i>et al.</i> [19]	-	-	48.1
Qi <i>et al.</i> [28]	57.0	-	45.9
Lee <i>et al.</i> [29]	60.1	81.5	47.7
Zhang <i>et al.</i> [25]	-	82.3	48.8
Chen <i>et al.</i> [33]	-	82.5	49.4
Hu <i>et al.</i> [45]	-	-	48.1
Zhou <i>et al.</i> [48]	60.5	82.5	49.3
Lin <i>et al.</i> [22]			49.5
Ours	61.44	87.62	49.57

D. Ablation Studies

To verify the effectiveness of different modules in the proposed BCINet, several leave-one-out experiments were conducted on the NYUv2 dataset. The BCIM, HPDC, and CGM were first removed from the BCINet and replaced by their corresponding simpler techniques for scene understanding as the Baseline method. Specifically, the BCIM was replaced by element-wise addition to fuse multimodal feature maps and the CGM was replaced by a combination of simple convolutional (1×1) + BN + ReLU layer with successive upsampling operations for final segmentation inference. Then, as listed in Table 3, many versions of BCINet were designed for comparing the performance of BCINet with different experimental settings.

Table 3 Ablation study of each component of BCINet on the NYUv2 dataset.

Methods	PixAcc	mAcc	mIoU
w/o HPDC	75.28	62.87	49.14
w/o CGM	76.14	66.29	51.53
w/o BCIM	76.66	66.46	52.05
Remove FEM in BCIM	76.37	66.29	51.54
Replace HPDC with ASPP	77.05	66.73	52.36
BCINet	77.17	66.78	52.95

1) *Effect of BCIM and FEM*: We designed the BCIM to explicitly model the complex relationships between RGB images and depth data by achieving bilateral cross-modal complementary information

interaction. When we removed the BCIM from the whole framework, as shown in Table III, the mIoU fell by 0.9%, verifying the effectiveness of our BCIM. Notably, if the FEM is removed from BCIM, the mIoU falls by 1.41%, which is worse than the results of removal of the whole BCIM. This indicates that inferior performance can result if we simply crossly fuse the original features of both modalities to each other without enhancing the feature correction. This substantiates our argument given in the Introduction and also validates the effectiveness of our FEM. For visualizations, as shown in the 4th column in Fig. 6, without BCIM, the model fails to obtain complementary cues, resulting in some objects with ambiguous appearance not being correctly classified. Additionally, owing to the lack of the FEM, the single modality discriminative features cannot be highlighted and extra noise is introduced, as can be seen in the 5th column. These situations are obviated when BCIM embedded with FEM is applied.

2) *Effect of HPDC*: As can be seen in Table III, the removal of HPDC causes a performance degradation of approximately 3.81%, which means that capturing diverse and anisotropic contexts plays a very important role in learning effective features. To verify the superiority of HPDC compared with the existing ASPP, we ablated HPDC by one more detailed experiment. As displayed in the 7th row of Table III, when we replace our HPDC with common ASPP, the performance declines by 0.58%, indicating that the three branches that adopt dilated rectangular convolutions in HPDC yield extra gains. As shown in the 4th column in Fig. 7, when HPDC is removed, the model fails to formulate long dependencies and large-scale variations between objects. Furthermore, if the HPDC is replaced by ASPP, objects with irregular shapes cannot be well segmented or correctly classified, which proves the effectiveness of the design of our HPDC.

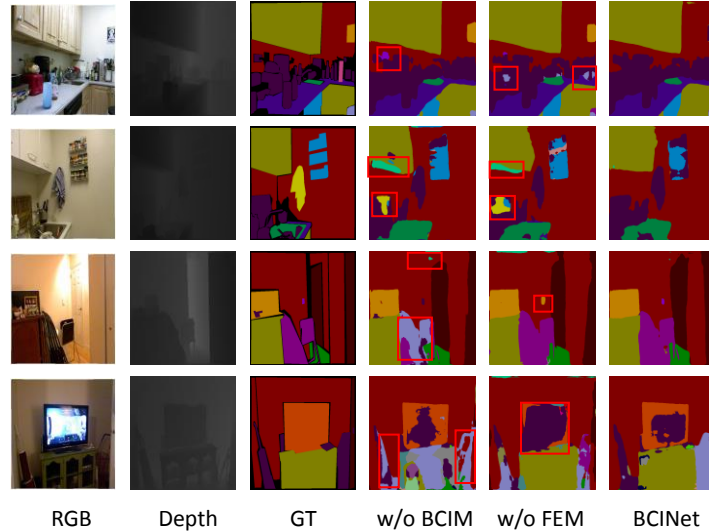


Fig. 6 Visualizations obtained by removing the BCIM and FEM.

3) *Effect of CGM*: It can be seen that the performance falls by 1.42% when CGM is not used. This indicates that CGM can better refine cross-modal high-level features by using the diverse contexts of

higher levels to guide the fusion of cross-modal lower-level feature maps in the encoder in a cascaded manner. As shown in the 4th column in Fig. 8, adding CGM results in better boundary delineation and higher accuracy for small objects.

Overall, Baseline + BCIM + HPDC + CGM yields the best performance of 52.95%, representing a 6.37% gain over the Baseline model. This implies that, with the collaboration of the BCIM, HPDC, and CGM, bilateral cross-modal complementary information is effectively obtained and diverse contexts are comprehensively exploited to improve the scene-understanding performance.

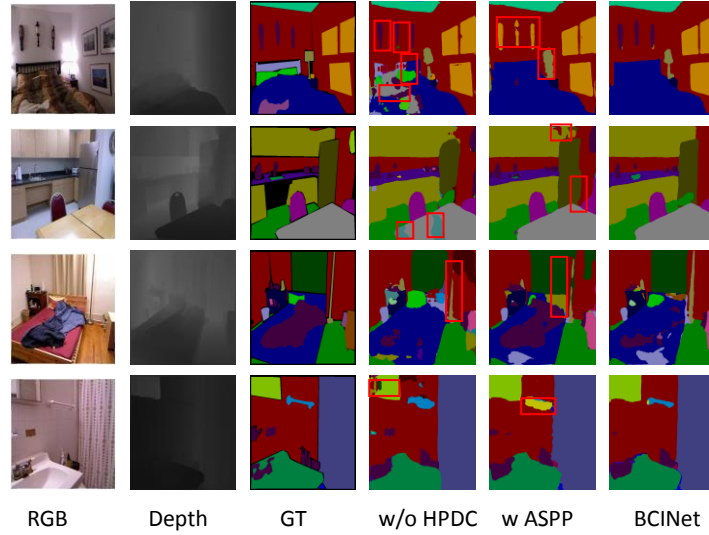


Fig. 7 Visualizations obtained by removing the HPDC and replacing HPDC with ASPP.

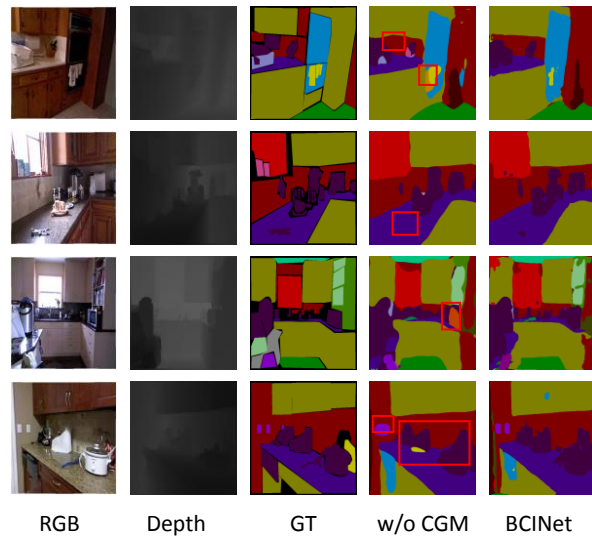


Fig. 8 Visualizations obtained by removing the CGM.

5. Conclusion

In this study, we proposed BCINet—which comprises BCIM, HPDC, and CGM—for RGB-D indoor

scene understanding. BCIM executes bilateral fusion strategies to achieve more effective cross-modal information interaction while preserving modality-specific features. HPDC is designed to capture more diverse and anisotropic contexts to adapt large-scale variations and objects with irregular shapes in indoor datasets. CGM guides the abundant and diverse high-level contexts captured by HPDC to aggregate them with lower-level features in the encoder, which facilitates boundary delineation and segmentation refinement. The results of experiments conducted on the NYUv2 and SUN datasets indicate that the proposed BCINet outperforms SOTA RGB-D scene-understanding models.

Acknowledgements

This work was supported by National Natural Science Foundation of China (61502429, 61672337, 61972357); the Zhejiang Provincial Natural Science Foundation of China (LY18F020012, LY17F020011); and Zhejiang Key R & D Program (2019C03135).

References

1. H. Xu, and J. Ma, EMFusion: An unsupervised enhanced medical image fusion network. *Information Fusion*, 76, (2021) 177–186.
2. K. Lin, Y. Li, J. Sun, D. Zhou, and Q. Zhang, Multi-sensor fusion for body sensor network in medical human–robot interaction scenario. *Information Fusion*, 57, (2020) 15–26.
3. H. Hermessi, O. Mourali, and E. Zagrouba, Multimodal medical image fusion review: Theoretical background and recent advances. *Signal Processing*, 183, (2021) 108036.
4. C. Hazirbas, L. Ma, C. Domokos, and D. Cremers, FuseNet: Incorporating depth into semantic segmentation via fusion-based CNN architecture, in *Proceedings of Asian Conference on Computer Vision (ACCV)*, 2016, pp. 213–228.
5. L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40, (2018) 834–848.
6. J. He, Z. Deng, L. Zhou, Y. Wang and Y. Qiao, Adaptive Pyramid Context Network for Semantic Segmentation, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 7511–7520.
7. J. Long, E. Shelhamer, and T. Darrell, Fully Convolutional Networks for Semantic Segmentation, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3431–3440.
8. H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, Pyramid Scene Parsing Network, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2881–2890.
9. C. Couprie, C. Farabet, L. Najman, and Y. Lecun, Indoor Semantic Segmentation using depth information, 2013, [Online] Available: <https://arxiv.org/pdf/1301.3572>.
10. H. Peng, B. Li, W. Xiong, W. Hu, and R. Ji, RGBD salient object detection: A benchmark and algorithms, in *Proceedings of the European Conference on Computer Vision(ECCV)*, 2014, pp. 92-109.
11. J. Han, H. Chen, N. Liu, C. Yan and X. Li, CNNs-Based RGB-D Saliency Detection via Cross-View Transfer and Multiview Fusion,

IEEE Transactions on Cybernetics, 48, (2018) 3171–3183.

12. Y. Cheng, R. Cai, Z. Li, X. Zhao and K. Huang, Locality-Sensitive Deconvolution Networks with Gated Fusion for RGB-D Indoor Semantic Segmentation, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1475–1483.
13. Z. Xiong, Y. Yuan, N. Guo and Q. Wang, Variational Context-Deformable ConvNets for Indoor Scene Parsing, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 3991–4001.
14. L. Ma, J. Stückler, C. Kerl and D. Cremers, Multi-view deep learning for consistent semantic mapping with RGB-D cameras, in Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2017, pp. 598–605.
15. H. Liu, W. Wu, X. Wang, and Y. Qian, RGB-D joint modelling with scene geometric information for indoor semantic segmentation, Multimedia Tools and Applications, 77, (2018) 22475–22488.
16. Y. He, W. Chiu, M. Keuper and M. Fritz, STD2P: RGBD Semantic Segmentation Using Spatio-Temporal Data-Driven Pooling, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 7158–7167.
17. M. Fayyaz, M. H. Saffar, M. Sabokrou, M. Fathy, F. Huang, and R. Klette, STFCN: Spatio-temporal fully convolutional neural network for semantic segmentation of street scenes, in Proceedings of Asian Conference on Computer Vision (ACCV), 2016, pp. 493–509.
18. F. Liu, G. Lin and C. Shen, Discriminative Training of Deep Fully Connected Continuous CRFs With Task-Specific Loss, IEEE Transactions on Image Processing, 26, (2017) 2127–2136.
19. D. Lin, G. Chen, D. Cohen-Or, P. Heng and H. Huang, Cascaded Feature Network for Semantic Segmentation of RGB-D Images, in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2017, pp. 1320–1328.
20. W. Wang and U. Neumann, Depth-aware CNN for RGB-D Segmentation, in Proceedings of the European Conference on Computer Vision (ECCV), 2020, pp. 135–150.
21. J. Yuan, W. Zhou and T. Luo, DMFNet: Deep Multi-Modal Fusion Network for RGB-D Indoor Scene Segmentation, IEEE Access, 7, (2019) 169350–169358.
22. D. Lin, R. Zhang, Y. Ji, P. Li and H. Huang, SCN: Switchable Context Network for Semantic Segmentation of RGB-D Images, IEEE Transactions on Cybernetics, 50, (2020) 1120–1131.
23. W. Zhou, J. Yuan, J. Lei and T. Luo, TSNet: Three-stream Self-attention Network for RGB-D Indoor Semantic Segmentation, IEEE Intelligent Systems, 36, (2021) 73–78.
24. J. Wang, Z. Wang, D. Tao, S. See, and G. Wang, Learning common and specific features for RGB-D semantic segmentation with deconvolutional networks, in Proceedings of the European Conference on Computer Vision (ECCV), 2016, pp. 664–679.
25. G. Zhang, J. -H. Xue, P. Xie, S. Yang and G. Wang, Non-Local Aggregation for RGB-D Semantic Segmentation, IEEE Signal Processing Letters, 28, (2021) 658–662.
26. Z. Li, Y. Gan, X. Liang, Y. Yu, H. Cheng, and L. Lin, LSTM-CF: Unifying context modeling and fusion with LSTMs for RGB-D scene labeling, in Proceedings of the European Conference on Computer Vision (ECCV), 2016, pp. 541–557.
27. J. Jiang, L. Zheng, F. Luo, and Z. Zhang, RedNet: Residual encoder-decoder network for indoor RGB-D semantic segmentation, 2018, [Online] Available: <https://arxiv.org/pdf/1806.01054>.
28. X. Qi, R. Liao, J. Jia, S. Fidler and R. Urtasun, 3D Graph Neural Networks for RGBD Semantic Segmentation, in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2017, pp. 5209–5218.
29. S. Lee, S. Park and K. Hong, RDFNet: RGB-D Multi-level Residual Feature Fusion for Indoor Semantic Segmentation, in

Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2017, pp. 4990–4999.

30. A. Valada, R. Mohan, and W. Burgard, Self-Supervised Model Adaptation for Multimodal Semantic Segmentation, *International Journal of Computer Vision*, 128, (2020) 1239–1285.
31. Z. Zhang, Z. Cui, C. Xu, Y. Yan, N. Sebe and J. Yang, Pattern-Affinitive Propagation Across Depth, Surface Normal and Semantic Segmentation, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 4101–4110.
32. D. Xu, W. Ouyang, X. Wang and N. Sebe, PAD-Net: Multi-tasks Guided Prediction-and-Distillation Network for Simultaneous Depth Estimation and Scene Parsing, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, 2018, pp. 675–684.
33. X. Chen, K. Lin, J. Wang, W. Wu, C. Qian, H. Li, and G. Zeng, Bi-directional Cross-Modality Feature Propagation with Separation-and-Aggregation Gate for RGB-D Semantic Segmentation, 2020, [Online] Available: <https://arxiv.org/pdf/2007.09183>.
34. H. Chen and Y. Li, Three-Stream Attention-Aware Network for RGB-D Salient Object Detection, *IEEE Transactions on Image Processing*, 28, (2019) 2825–2835.
35. X. Zhang, T. Wang, J. Qi, H. Lu and G. Wang, Progressive Attention Guided Recurrent Network for Salient Object Detection, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 714–722.
36. N. Liu, J. Han and M. Yang, PiCANet: Learning Pixel-Wise Contextual Attention for Saliency Detection, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 3089–3098.
37. J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang and H. Lu, Dual Attention Network for Scene Segmentation, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 3141–3149.
38. Y. Yuan and J. Wang, OCNet: Object context network for scene parsing, 2018, [Online] Available: <https://arxiv.org/pdf/1809.00916>.
39. J. Jia, PSANet: Point-wise Spatial Attention Network for Scene Parsing, in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 267–283.
40. X. Wang, R. Girshick, A. Gupta, and K. He, Non-local Neural Networks, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7794–7803.
41. Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei and W. Liu, CCNet: Criss-Cross Attention for Semantic Segmentation, in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 603–612.
42. Y. Cao, J. Xu, S. Lin, F. Wei and H. Hu, GCNet: Non-Local Networks Meet Squeeze-Excitation Networks and Beyond, in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, 2019, pp. 1971–1980.
43. J. Hu, L. Shen, S. Albanie, G. Sun and E. Wu, Squeeze-and-Excitation Networks, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7132–7141.
44. Z. Zhou, Z. Wang, H. Lu, S. Wang, and M. Sun, Multi-Type Self-Attention Guided Degraded Saliency Detection, in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, pp. 13082–13089.
45. X. Hu, K. Yang, L. Fei and K. Wang, ACNET: Attention Based Network to Exploit Complementary Features for RGBD Semantic Segmentation, in *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, 2019, pp. 1440–1444.
46. L. Deng, M. Yang, T. Li, Y. He and C. Wang, RFBNet: deep multimodal networks with residual fusion blocks for RGB-D semantic segmentation, 2019, [online] Available:<https://arxiv.org/pdf/1907.00135>.

47. M. Zhang, Y. Zhang, Y. Piao, B. Hu and H. Lu, Feature Reintegration over Differential Treatment: A Top-down and Adaptive Fusion Network for RGB-D Salient Object Detection, in Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 4107–4115.
48. H. Zhou, L. Qi, Z. Wan, H. Huang and X. Yang, RGB-D Co-attention Network for Semantic Segmentation, in Proceedings of Asian Conference on Computer Vision (ACCV), 2020, pp. 519–536.
49. R. Vemulapalli, O. Tuzel, M. Liu and R. Chellappa, Gaussian Conditional Random Field Network for Semantic Segmentation, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 3224–3233.
50. S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang and P. H. S. Torr, Conditional Random Fields as Recurrent Neural Networks, in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2015, pp. 1529–1537.
51. H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, Pyramid scene parsing network, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2881–2890.
52. H. Mei, G. Ji, Z. Wei, X. Yang, X. Wei, and D. Fan, Camouflaged object segmentation with distraction mining, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 8772–8781.
53. J. Zhang, C. Long, Y. Wang, X. Yang, H. Mei, and B. Yin, Multicontext and enhanced reconstruction network for single image super resolution, in Proceedings of the IEEE International Conference on Multimedia and Expo (ICME), 2020, pp. 1–6.
54. K. He, X. Zhang, S. Ren and J. Sun, Deep Residual Learning for Image Recognition, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.
55. M. Yang, K. Yu, C. Zhang, Z. Li, and K. Yang, DenseASPP for semantic segmentation in street scenes, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3684–3692.
56. G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, Densely connected convolutional networks, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4700–4708.
57. L. Wang, R. Chen, L. Zhu, H. Xie and X. Li, Deep Sub-Region Network for Salient Object Detection, IEEE Transactions on Circuits and Systems for Video Technology, 31 (2021) 728–741.
58. N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, Indoor segmentation and support inference from RGBD images, in Proceedings of the European Conference on Computer Vision (ECCV), 2012, pp. 746–760.
59. S. Song, S. P. Lichtenberg and J. Xiao, SUN RGB-D: A RGB-D scene understanding benchmark suite, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 567–576.
60. J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei, Imagenet: A large-scale hierarchical image database, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2009, pp. 248–255.
61. W. Liu, A. Rabinovich, and A. C. Berg, Parsenet: Looking wider to see better, 2015, [online] Available: <https://arxiv.org/pdf/1506.04579>