CFANet: The Cross-Modal Fusion Attention Network for Indoor RGB-D Semantic Segmentation

Long-Fei Wu , Dan Wei * , Chang-An Xu

Posted Date: 21 April 2025

doi: 10.20944/preprints202504.1682.v1

Keywords: cross-modal fusion; RGB-D; feature extraction; feature interaction

 doi:10.20944/preprints202504.1682.v1

Article

CFANet: The Cross-Modal Fusion Attention Network for Indoor RGB-D Semantic Segmentation

Long-Fei Wu [1], Wei Dan [1,*] and Chang-An Xu [2]

[1] School of Mechanical and Automotive Engineering, Shanghai University of Engineering Science, Shanghai, 201620, China
[2] South China Agricultural University, Guangzhou 510642, China
* weiweidandan@163.com

Abstract: The indoor image semantic segmentation technology is applied to fields such as smart homes and indoor security. The challenges faced by semantic segmentation techniques using RGB images and depth maps as data sources include the semantic gap between RGB images and depth maps and the loss of detailed information. To address these issues, a multi-head self-attention mechanism is adopted to adaptively align features of the two modalities and perform feature fusion in both spatial and channel dimensions. Appropriate feature extraction methods are designed according to the different characteristics of RGB images and depth maps. For RGB images, asymmetric convolution is introduced to capture features in horizontal and vertical directions, enhance short-range information dependence, mitigate the gridding effect of dilated convolution, and introduce Criss-Cross Attention to obtain contextual information from global dependency relationships. On the depth map, a strategy of extracting significant unimodal features from the channel and spatial dimensions is used. A lightweight skip connection module is designed to fuse low-level and high-level features. In addition, since the first layer contains the richest detailed information and the last layer contains rich semantic information, a feature refinement head is designed to fuse the two. The method achieves mIoU of 53.86% and 51.85% on the NYUDv2 and SUN-RGBD datasets, which is superior to mainstream methods.

Keywords: cross-modal fusion; RGB-D; feature extraction; feature interaction

1. Introduction

Semantic segmentation for indoor scenario is a dense prediction task that aims to assign a category label to each pixel in an image. It is applied in medical image analysis, industrial robots, intelligent driving and other fields[1]. In order to improve the accuracy of semantic segmentation, many researchers use convolutional neural Network (CNN) [2] to extract image features. The powerful feature extraction ability of CNN and recombination of multilevel information have led to notable advances in segmentation. For example, He et al. [3] pro-posed a semantic segmentation model based on pyramid scene parsing, which is characterized by combining kernels of different sizes to create a spatial pyramid pooling network. PointFlow [4] is proposed by Huang et al, which adaptively uses the high-semantic low-resolution feature map to enhance the low-semantic high-resolution feature map to obtain the high-semantic high-resolution feature map. Although CNN shows strong performance in information representation, RGB images are planarization of 3D space and lose depth information. Only using RGB images as a single data source, deep learning networks cannot learn enough information in com-plex and diverse scenes.

RGB images may generate noise due to similar texture characteristics among different objects. However, depth maps can provide relative distance information of objects, unaffected by the color and texture of similar objects, and can also distinguish the relative positions of occluded objects. The information provided by depth maps can compensate for the shortcomings of indoor RGB images, such as occlusions and similar textures. As shown in Figure 1, depending on the timing of the fusion

 doi:10.20944/preprints202504.1682.v1

of RGB and depth map features within the network, the fusion methods can be categorized into three types: 1) Early fusion: RGB and depth maps are concatenated first and then passed through convolutional layers to generate feature maps; 2) Late fusion: The network adopts a dual-branch structure, where these two branches independently extract corresponding RGB and depth features, and finally, the extracted features are fused; 3) Multi-level fusion: This method gradually fuses the features of RGB and depth maps layer by layer, enhancing the complementarity of the two through dynamic weight configuration. Due to the significant semantic gap between RGB and depth maps, the semantic segmentation results using early fusion are not ideal. Although the methods in (b)-(c) have made some progress, there is still room for exploration of the correlation and interaction between RGB and depth data. In addition, depth maps may also contain noise. As shown in Figure 2(a), due to the limitations of camera hardware, when objects are far from the camera, their boundaries are not clear, which can easily introduce noise during boundary information extraction. As shown in Figure 2(b), different objects at the same distance from the camera may be segmented as a single object.

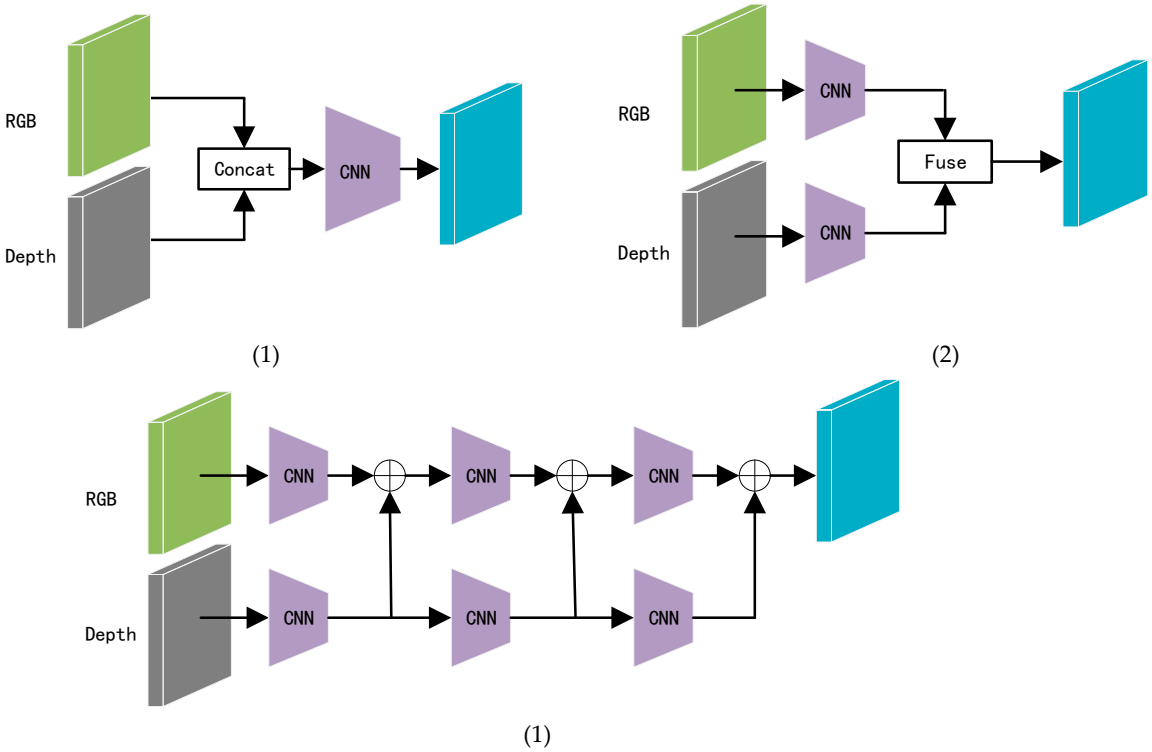
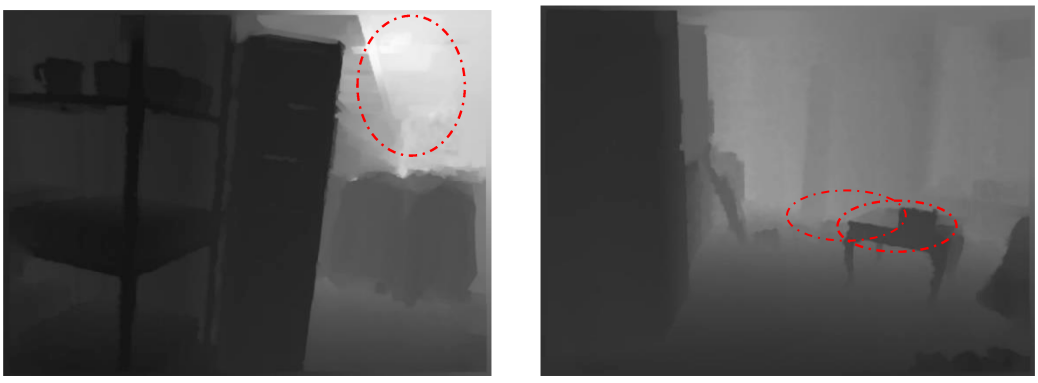


Figure 1. Fusion methods. (1) Early fusion. (2) Late fusion. (3) Multi-level fusion.



(a) (b)

Figure 2. Noise in depth maps. (a) Type one noise: The boundary in the distance is not clear.; (b) Type two noise: different objects at the same distance from the camera may be segmented as a single object.

 doi:10.20944/preprints202504.1682.v1

Many studies [5,6] employ similar methods to extract features from two modalities of data (RGB and depth maps) or simply use depth maps as a supplement to RGB images, ignoring the differences and complementarity of these two modalities. RGB images in indoor environments contain rich textures and boundary details. However, challenges such as occlusions and similar textures between different objects still exist. While depth maps pro-vide information about the relative positions of objects, they offer limited effective boundary information. Considering the modal differences between RGB and depth maps, we have designed two different feature extraction modules: the RGB Feature Extraction Module (BFEM) and the Depth Feature Extraction Module (DFEM). BFEM adds asymmetric convolutions in ASPP [7] to capture features in horizontal and vertical spatial directions and introduces Criss-Cross Attention to obtain contextual information in global dependencies; DFEM extracts significant unimodal features of depth maps from both channel and spatial dimensions. However, there is an obvious semantic gap between the data of these two modalities. Therefore, an Adaptive Feature Complementary Fusion Module (AFFCM) is proposed to automatically align the features of these two modalities.

As convolutional neural networks deepen, there is a possibility of losing information pertaining to small-scale objects. Multilevel contextual information plays a significant role in the segmentation of small-scale objects. The rich details of object boundaries are encompassed within low-level contextual information, while the relationships between different objects are encompassed within high-level contextual information. Researchers have conduct-ed extensive studies to fully capture the characteristics of multi-scale objects present in images. Many classical network models have been proposed for this purpose, such as the image pyramid [8], feature pyramid [9], and spatial pyramid pooling [3]. Detailed information is crucial for capturing multi-scale object details. Low-level features provide essential details such as textures, spatial relationships, and other key information for multi-scale objects.

The feature map output by the first layer of the encoder preserves detailed information but lacks semantic content. Conversely, the feature map output by the last layer of the decoder contains rich semantic information, but it loses detailed information as the network deepens. In previous studies, a common approach was to perform an additive operation on these two feature maps. While this method reduces computational complexity, it often leads to a decrease in semantic segmentation accuracy. In this paper, we design a Feature Refinement Head (FRH) that refines feature maps from both spatial and channel perspectives. Additionally, we introduce a novel Skip Connection Module (SCM) that fuses multi-scale feature maps to enhance the network's robustness towards multi-scale objects.

The main contributions of this paper are as follows:

(1) We introduce a novel dual-branch RGB-D semantic segmentation network named CFANet. Tailored feature extraction modules are designed based on the distinct characteristics of RGB and depth maps, subsequently enhancing segmentation accuracy through adaptive cross-modal feature fusion.
(2) BFEM introduces asymmetric convolution on the basis of dilated convolution to alleviate the gridding effect of dilated convolution. Additionally, it achieves rich contextual learning through dense connections and Criss-Cross Attention. DFEM extracts significant unimodal features from both the channel and spatial dimensions of depth maps.
(3) Guide the RGB branch and depth map branch to interact rather than simply treating the depth map as a complement to the RGB image.
(4) AFFCM employs a multi-head self-attention mechanism to address the semantic discrepancies between RGB and depth map features, resulting in their adaptive alignment. This process effectively enhances the complementary information exchange between the two modalities and mitigates redundancy.
(5) We adopt different strategies for feature maps of different scales. Multi-scale feature maps are fused through SCM; considering that the first layer contains the richest detailed information and the last layer contains the richest semantic information, we designed FRH to fuse both.

 doi:10.20944/preprints202504.1682.v1

2. Related Works

2.1. RGB-D Semantic Segmentation

Due to the intricate occlusion relationships, diverse object sizes, and high similarity in texture and color characteristics within indoor scenes, information provided by a single RGB image is insufficient for indoor semantic segmentation tasks. With the advancement of depth cameras, researchers have incorporated depth maps as an additional data source for image semantic segmentation. For instance, Kazerouni et al. [10] proposed an RGB-D image semantic segmentation network based on the fusion of multimodal image features. This method employs a direct summation approach to progressively merge RGB image features with depth images at different hierarchical levels. Zheng et al. [11] introduced a depth-similar convolution, replacing regular convolutional layers and average pooling layers in the encoder. Zhou et al. [12] presented FRNet, a method for indoor scene understanding, which extracts features hierarchically in a top-down manner. However, the aforementioned methods simply fuse depth maps and RGB images through a straightforward summation process.

To enhance the fusion between RGB images and depth maps, Zhou et al. [13] applied a self-attention module during the decoding stage to adaptively merge depth map features and RGB features. Cao et al. [14] introduced a convolution named ShapeConv, which can learn an adaptive balance between shape information and the preservation of base information, allowing the network to focus more on shape information when necessary, thereby benefiting the RGB-D semantic segmentation task. Zhang et al. [15] proposed a non-local aggregation method for RGB-D semantic segmentation, integrating features from both spatial and channel perspectives to achieve multi-level information fusion between RGB images and depth maps. Zhou et al. [16] introduced a novel feature enhancement method aimed at improving the features extracted from RGB and depth maps.

The aforementioned approach has made significant contributions to feature extraction and fusion. Building upon prior research, CFANet takes into full consideration the distinct characteristics of depth maps and RGB im-ages. It employs appropriate modules to extract features from these two modalities separately and ensures an interaction between the RGB branch and depth map branch, avoiding a simplistic view of the depth map merely as a complement to the RGB image.

2.2. Attention Mechanisms

Attention mechanisms aim to mimic the human visual selective attention capability by assigning weights based on the importance of different parts in a feature map. The attention mechanisms proposed by scholars can be categorized into four types: spatial attention, channel attention, spatial and channel mixed attention, and self-attention.

Spatial attention enhances the network's ability to capture important objects in the feature map by applying weighted operations across different regions in the spatial dimension. Mnih et al. [17] perform multi-level attention focusing at different scales through simultaneous processing of sequence and spatial data, showcasing strong adaptability and flexibility at the expense of higher computational complexity. Jaderberg et al. [18] address the issue of input images being non-amenable to geometric transformations by designing a parameterized spatial transformer sub-network, converting the geometric rectification process into a differentiable operation chain, enabling the model to learn the optimal sample spatial registration strategy in an end-to-end manner. Huang et al. [19] mitigate the high computational complexity and resource consumption issues associated with traditional non-local methods by introducing Criss-Cross Attention. This mechanism captures correlated features along both horizontal and vertical directions, computing the correlation only between the target pixel and its row and column positions with significant spatial correlation, and integrates long-range dependency information through an adaptive weighted fusion strategy.

Channel attention aims to address the issue of feature channel dependency by adaptively assigning weights to channels, thereby enhancing useful features. Hu et al. [20] compress the feature map into a multidimensional vector, assign weights to the parameters in the vector through a loss function, and dynamically adjust the original feature map based on their importance. Qin et al. [21] extend the global average pooling operation in the channel attention mechanism to a two-

 doi:10.20944/preprints202504.1682.v1

dimensional discrete cosine transform (DCT) form, demonstrating that global average pooling is a special case of two-dimensional DCT. This introduces more frequency components to fully extract feature information. Wang et al. [22] believe that the dimensionality reduction operation in [20] has negative impacts. They propose an Efficient Channel Attention (ECA) module. ECA uses one-dimensional convolutions to acquire information from adjacent channels, yielding more precise channel attention information. ECA can adaptively adjust the size of the one-dimensional convolution in the module, allowing high-dimensional channels to have longer-range interactions and low-dimensional channels to have shorter-range interactions. Yang et al. [23] combine gating mechanisms and normalization methods to selectively model different channel feature information.

Spatial and channel mixed attention adaptively selects important channels and spatial regions, weighting and summing the results of channel attention and spatial attention to obtain a blended feature vector. Woo et al.'s [24] proposed Convolutional block attention module (CBAM) consists of two parts: the Channel Attention Module (CAM) focuses on the importance of each channel, capturing significant features in the image, while the Spatial Attention Module (SAM) selects for each spatial position, capturing meaningful local regions in the image. Hu et al. [25], utilizing the concept of feature separation and interaction, explore relationships between image features, but encounter the issue of information loss during information exchange and recombination. Fu et al.[26], employing a spatial pyramid attention mechanism, effectively handle feature information at different scales, enhancing the model's representational capacity. However, this model has a higher computational complexity, resulting in substantial training and tuning costs, requiring significant computational resources and time investment.

Self-Attention was initially applied in the field of Natural Language Processing (NLP) due to its powerful ability to model global contextual information. Dosovitskiy et al. [27] split input vectors into designated heads and independently compute self-attention for each head. Each head learns different features independently, complementing and collaborating with each other, thereby enhancing the model's expressive capacity. Liu et al.[28], to ad-dress the computational complexity and loss of inter-token correlation information when dealing with high-resolution images in self-attention mechanisms, introduce a sliding window mechanism. The sliding window multi-head self-attention layer segments the input tensor into non-overlapping blocks and performs multi-head self-attention calculations on each block. Dong et al. [29] propose the innovative Cross Sliding Window Self-Attention (CSWin Transformer), which adopts a cross-shaped window self-attention mechanism (CW-MSA) to replace the W-MSA and SW-MSA modules in Swin Transformer. This enables local modeling of input features, preserving more detailed information and fully utilizing global features, significantly improving model accuracy without increasing computational complexity.

Building upon these outstanding research contributions, we constructed CFANet. In CFANet, we integrated Criss-Cross Attention [19] into BFEM. DFEM utilizes attention mechanisms to extract significant unimodal features from the spatial and channel dimensions of the depth map. AFCFM employs Multi-Head Attention to align RGB features and depth map features. Drawing inspiration from attention mechanisms, it enhances the complementary information between the two modalities and diminishes redundant information.

3. Method

3.1. Overview of the Architecture

Figure 3 illustrates the framework of CFANet proposed in this paper, which adopts the classical encoder-decoder architecture to achieve end-to-end semantic segmentation. During the encoding phase, RGB images and depth maps are input separately into their respective network branches. In both branches, we employ ResNet50[20] to extract features from the images, yielding four feature maps at different resolutions for each type of image. L_i^B (i=1,2,3,4) represent the extracted RGB image features, and L_i^D represent the extracted depth image features. BFEM is designed to develop RGB image features enriched with contextual information, while DFEM focuses on capturing significant unimodal features from the depth map. Applying BFEM to L_i^B results in the re-extracted feature maps R_i^B

 doi:10.20944/preprints202504.1682.v1

. The addition of R_i^B and L_i^D produces new feature maps L_{i+1}^D, and the addition of R_i^D and L_i^B produces a new feature map L_{i+1}^B, facilitating feature interaction. This process is illustrated by the following formula (1):

$$\begin{cases} L_{i+1}^D = L_i^D + R_i^B \\ L_{i+1}^B = L_i^B + R_i^D \end{cases}, (i = 1, 2, 3) \qquad (1)$$

As illustrated in Figure 3, to mitigate the semantic gap present in different modal images, AFCFM is employed to merge feature maps from different modalities, resulting in the fused features H_i, as described in the following text. SCM designed in this paper strikes a good balance between accuracy and computational efficiency. Considering the unique characteristics of E_1 and L_1^B, we have developed the Feature Refinement Head (FRH). The FRH further refines the feature maps from both spatial and channel perspectives.

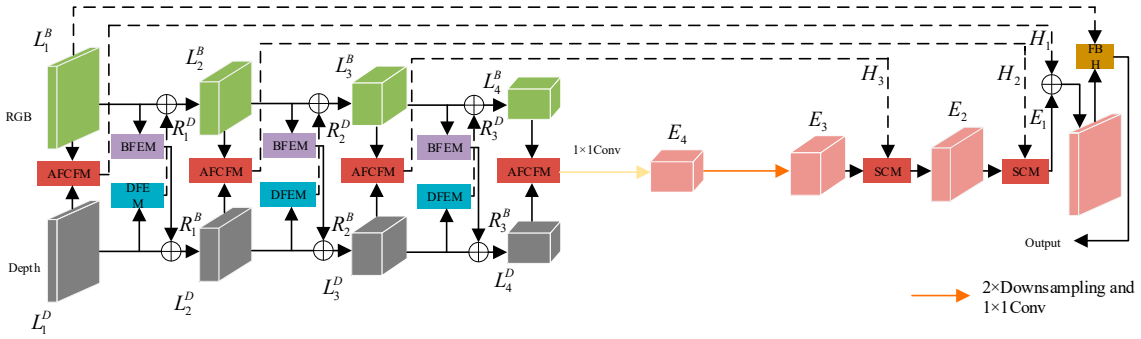


Figure 3. Overall architecture of CFANet

3.2. RGB Feature Extraction Module and Depth Map Feature Extraction Module

The modal differences between RGB images and depth maps are primarily manifested in the way information is expressed. RGB images offer rich information about object color, texture, and appearance, while depth maps provide information about the distance or depth of objects, allowing for a better description of spatial relation-ships and shape information between objects in the scene. Because these two modalities provide distinct information, effectively leveraging both becomes a crucial challenge in RGB-D semantic segmentation tasks. To ad-dress this challenge, we have tailor-made two feature extraction modules to suit the characteristics of different modal data.

3.2.1. RGB Feature Extraction Module

Indoor RGB images exhibit the following characteristics: (1) Indoor scenes comprise objects of various scales, such as small-sized decorations and large-sized wardrobes. (2) Indoor image backgrounds are complex, potentially containing numerous background objects similar to the target objects. To address these challenges, there is a need to increase the receptive field and learn rich contextual information. In previous research, dilated convolutions were introduced to enlarge the receptive field without introducing a large number of parameters. However, dilated convolutions may lead to a loss of correlation between adjacent pixels. The BFEM not only alleviates the aforementioned challenges but also enables the learning of abundant contextual information.

To mitigate the gridding issue associated with dilated convolutions, BFEM incorporates asymmetric convolutions [30]. The structure of BFEM is illustrated in Figure 4. After filtering the feature map L_i^B with dilated convolutions, we immediately fuse it with the feature map processed by the corresponding asymmetric convolutions. Additionally, we sum up the previously processed feature maps to enhance dense connections, as shown in Figure 4(a). Dense connections allow for the repeated reuse of signals, providing substantial expansion of the receptive field and enabling the learning of more contextual information [22]. We implement the above processes using the following formula (2):

 doi:10.20944/preprints202504.1682.v1

$$\begin{cases} F_i^1 = \delta\left[\left(L_i^B \otimes K_i^1\right) + C_1\left(L_i^B\right)\right] \\ F_i^2 = \delta\left[\left(F_i^1 \otimes K_i^2\right) + C_2\left(L_i^B\right)\right] \\ F_i^3 = \delta\left[\left(F_i^1 \otimes K_i^3\right) + \left(F_i^2 \otimes K_i^3\right) + C_3\left(L_i^B\right)\right] \end{cases} \quad (2)$$

In the formula (2), F_i^{m} ($m = 1,2,3$; $i = 1,2,3$)represents the feature map generated after each densely connected operation; K_i^1 represents the dilated convolution with a dilation rate of 1, K_i^2 represents the dilated convolution with a dilation rate of 2, and K_i^3 represents the dilated convolution with a dilation rate of 3; C_1 represents a 3×1 convolution followed by a 1×3 convolution, C_2 represents a 4×1 convolution followed by a 1×4 convolution, and C_3 represents a 5×1 convolution followed by a 1×5 convolution; F_i^B $(i = 1,2,3)$ represents the feature map generated by the i-th layer in the RGB network branch; δ is a 2×downsampling followed by a 1×1 convolution.

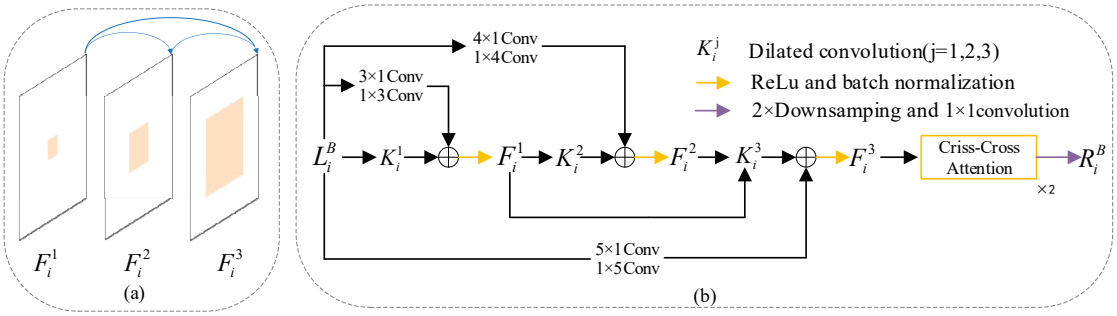


Figure 4. Detailed structure of the proposed BFEM. (a) Diagram of dense connection; (b) Flowchart of BFEM.

In order to effectively capture global contextual information while mitigating computational overhead, we employed cross-cross attention [19]. The feature map undergoes two cross-cross attention operations to yield F_i^3 ,as shown in Figure 4(b).

3.2.2. Depth Map Feature Extraction Module

Indoor depth maps may contain noise, especially in regions with uneven lighting or texture absence. Consequently, features provided by the depth map may introduce information detrimental to segmentation accuracy. To alleviate this issue, we employ DFEM to extract significant unimodal features in both spatial and channel dimensions.

The structure of DFEM is illustrated in Figure 5. The two upper branches extract weighted features of L_i^D from the spatial dimension, while the two lower branches extract weighted features of L_i^D from the channel dimension. Specifically, the two upper branches take the feature maps processed by global max-pooling and global average-pooling, input them into a 3×3 convolution for adaptive parameter updates, apply a sigmoid function to obtain spatial feature weights, and finally perform element-wise multiplication of the spatial feature weights with the input feature L_i^D to obtain spatial-weighted features. Similarly, the two lower branches take the feature maps processed by channel max-pooling and channel average-pooling, input them into a 3×3 convolution for adaptive parameter updates, apply a sigmoid function to obtain channel feature weights, and finally perform element-wise multiplication of the channel feature weights with the input feature L_i^D to obtain channel-weighted features. The spatial-weighted features and channel-weighted features are added, and after 2×downsampling followed by 1×1 convolution to adjust the channel number, the result R_i^D is obtained. This process is illustrated by the following formula (3):

$$R_i^D = \delta\left\{\sum_{P=1}^{4}\left[\alpha\left(\mathrm{Conv}(P(L_i^D))\right) \otimes L_i^D\right]\right\} \quad (3)$$

In the formula (3), when $P = 1$, P represents global max pooling; when $P = 2$, P represents global average pooling; when $P = 3$, P represents channel max pooling; and when $P = 4$, P represents channel

 doi:10.20944/preprints202504.1682.v1

average pooling. ⊗ denotes element-wise multiplication, *C*onv signifies a 3×3 convolution. α is the sigmoid function, and δ represents a 2×downsampling followed by 1×1 convolution.

3.3. Adaptive Feature Complementary Fusion

The feature maps extracted from RGB and depth images not only contain complementary information but al-so entail redundant details. Striking a balance between enhancing complementary information and reducing redundancy is our objective. Leveraging the potent feature representation capabilities of the Multi-Head Attention [31], which empowers the model to learn relationships between different parts and representations effectively. AFCFM utilizes a Multi-Head Attention mechanism to adaptively align the feature maps generated from RGB and depth images. Subsequently, it enhances the representation of complementary information from both channel and spatial perspectives, mitigating redundancy.

As illustrated in Figure 6, the feature map R_i^B obtained from the RGB image branch and the feature map R_i^D obtained from the depth image branch are linearly mapped to derive Q, K and V. It is noteworthy that, to align R_i^B and R_i^D, this paper employs Q as one of the inputs for the Multi-Head Attention mechanism between them. This process can be described by the following formula (4):

$$\begin{cases} R_i^{RB} = MHSA(Q^D, K^B, V^B) \\ R_i^{RD} = MHSA(Q^B, K^D, V^D) \end{cases} \quad (4)$$

In the formula (4), *MHSA* represents the Multi-Head Attention. Q^B, K^B, V^B is the vector obtained from the RGB image, and Q^D, K^D, V^D is the vector obtained from the depth image. R_i^{RB} represents the aligned RGB feature map, and R_i^{RD} represents the aligned depth feature map after the alignment process.

Influenced by the idea of the attention mechanism, we assign different weights to R_i^{RB} and R_i^{RD} to enhance complementary information while reducing redundancy. In specific experiments, we adopt the framework of CBAM [24] and apply the Attention mechanism separately in the channel and spatial dimensions. However, the original Channel Attention mechanism in CBAM employs two consecutive fully connected layers, and the dimension reduction operation in it has a negative impact on the prediction of Channel Attention, resulting in lower computational efficiency. We replace this with a simple one-dimensional convolution to adaptively explore the weights for each channel. The Channel Attention Module's structure is represented as CA in Figure 6, and the structure of Spatial Attention is denoted as SA. The weight information obtained from CA and SA is then mapped to R_i^{RB} and R_i^{RD}, and the fused feature H_{i+1} is derived through addition. This process can be described by the following formula (5) and formula (6):

$$R_i^{CON} = Concat(R_i^{RB}, R_i^{RD}) \quad (5)$$

$$H_{i+1} = \alpha\left[SA(R_i^{CON}) + CA(R_i^{CON})\right] \otimes R_i^{RB} + \alpha\left[SA(R_i^{CON}) + CA(R_i^{CON})\right] \otimes R_i^{RD} \quad (6)$$

In the formulas, $E_{i-1} = PS\{\alpha[Conv(Con(E_i, H_i))] \otimes Con(E_i, H_i)\}$ represents Spatial Attention, *CA* represents Channel Attention. α is the sigmoid function, ⊗ is element-wise multiplication. H_{i+1} is the fused feature map obtained after the attention mechanisms..

 doi:10.20944/preprints202504.1682.v1

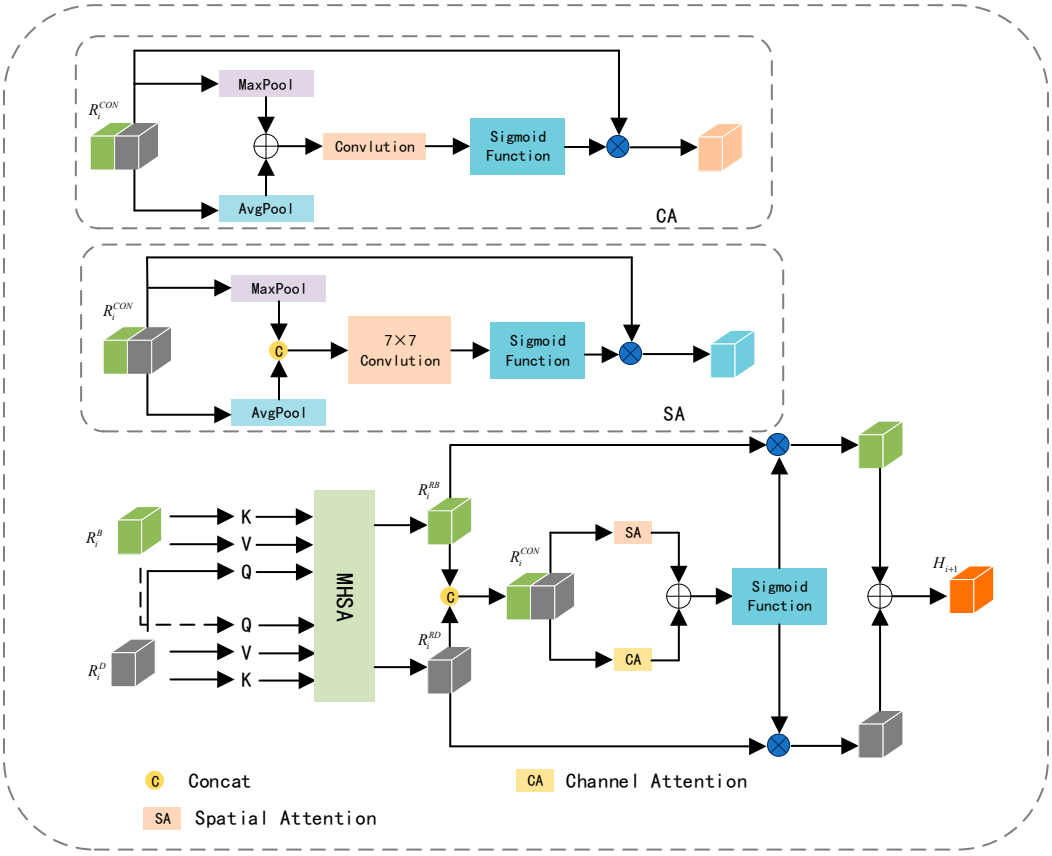


Figure 6. Detailed structure of the proposed AFCFM

3.4. Skip Connection Module and Feature Refinement Head

The feature maps generated by the encoder retain rich detailed information but lack semantic content Conversely, those generated by the decoder contain abundant semantic information but suffer from significant in-formation loss as the network deepens. In previous studies, a common approach has been to perform a simple addition operation on these two feature maps. While computationally less intensive, this approach compromises semantic segmentation accuracy. To address this, we propose two efficient modules, namely SCM and RFH.

3.4.1. Feature Refinement Head

The structure of FRH is illustrated in Figure 7(a). FRH module aims to deeply fuse L_1^B and E_1, maximizing the utilization of both detailed and semantic information. Building upon feature fusion, the feature maps undergo refinement through spatial and channel branches. Specifically, the spatial branch employs 3×3 depthwise separable convolutions to generate attention maps, aiming to reduce parameter count and enhance computational efficiency. The channel branch, on the other hand, first compresses the feature maps into a one-dimensional vector using global average pooling. It then utilizes 1×1 convolutions to reduce the number of channels to one-fourth of the original, followed by another 1×1 convolution to restore it to the original channel count. This operation is in-tended to increase network depth while mitigating the risk of overfitting. Ultimately, the feature maps generated by the spatial and channel paths are further fused through summation. Additionally, to prevent network degradation during training, this paper introduces a residual connection mechanism, enhancing the stability and effectiveness of network training.

3.4.2. Skip Connection Module

In previous studies, the common practice was to use a simple addition operation on feature maps or employ complex modules to fuse feature maps from the decoding and encoding stages. Unlike prior research, we utilize Concatenation to combine the feature maps input to SCM, obtaining features with higher representational capacity. In contrast to more complex models, we use only a 1×1

 doi:10.20944/preprints202504.1682.v1

convolution to update parameters. Additionally, SCM employs Pixel Shuffle [32] to reorganize channels, enabling an increase in feature map resolution and retention of more detailed information without introducing new parameters. The structure of SCM is illustrated in Figure 7(b). This process can be described by the following formula (7):

$$E_{i-1} = PS\{\alpha[Conv(Con(E_i, H_i))] \otimes Con(E_i, H_i)\} \quad (7)$$

Where *PS* denotes Pixel Shuffle, *Con* represents Concat operation, α is the sigmoid function, $\otimes$ is element-wise multiplication. *Conv* signifies a 1×1 convolution.

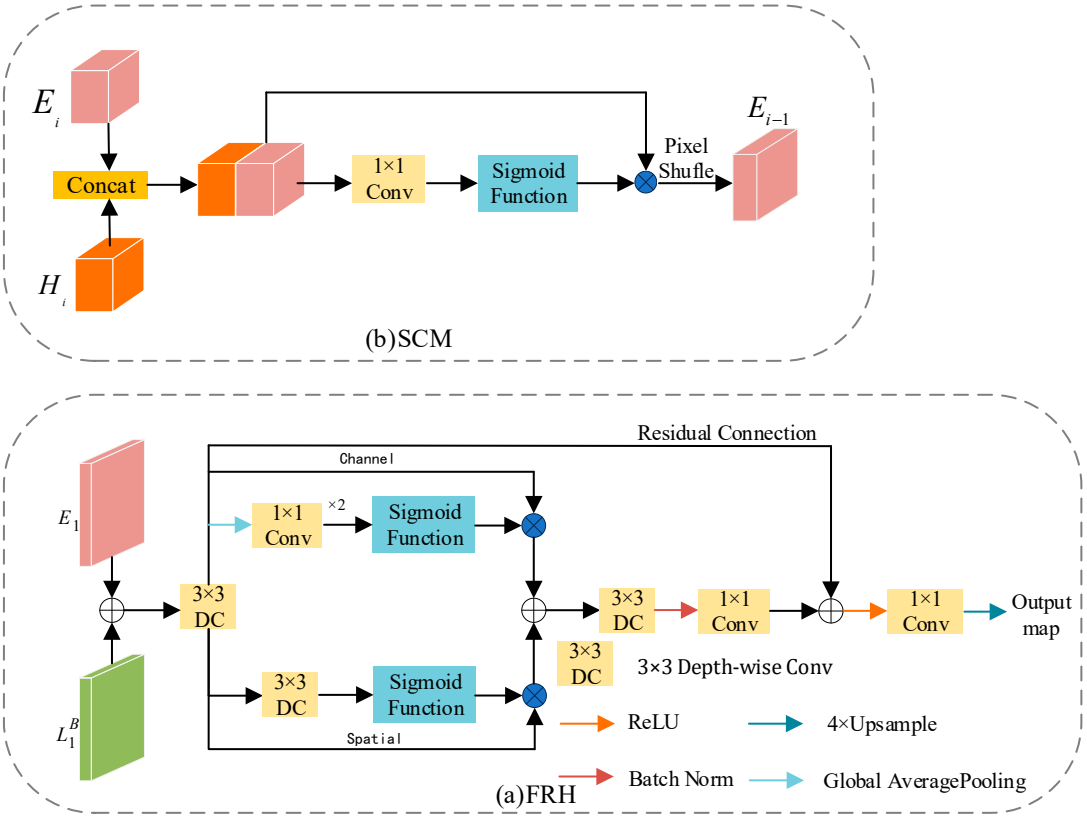


Figure 7. Detailed structure of the proposed FRH and SCM. (a) Architectural diagram of FRH; (b) Architectural diagram of SCM.

4. Experiment

4.1. Dataset and Metrics

The SUN-RGBD dataset is a highly comprehensive dataset for indoor scene understanding, comprising 10,335 pairs of images depicting various indoor scenes such as living rooms, bedrooms, kitchens, bathrooms, and more. A total of 37 different object categories are annotated in this dataset. Among the images, 5,285 are designated for training, while 5,050 images are reserved for testing purposes.

The NYUDv2 dataset is composed of video sequences capturing various indoor scenes recorded by Microsoft Kinect's RGB and Depth cameras. This dataset comprises 1,449 pairs of densely annotated RGB and depth im-ages, covering a total of 40 different object categories. These categories encompass various objects commonly found in indoor scenes, including chairs, tables, TVs, people, and more. The dataset is split into a training set with 795 images and a test set containing the remaining 654 images.

We measured the proposed CFANet with state-of-the-art methods according to three measures: mean accuracy (mAcc.), pixel accuracy (PixAcc.), and mean intersection over union (mIoU).

 doi:10.20944/preprints202504.1682.v1

4.2. Implementation Details

CFANet is implemented using the PyTorch framework on 2 NVIDIA GeForce RTX 4080 GPUs. A ResNet50 pretrained on ImageNet is employed as the backbone network for CFANet. During the training phase, all input images are uniformly cropped to a resolution of 480×640 pixels. To enhance the model's generalization ability, data augmentation techniques such as random scaling, random cropping, random horizontal flipping, random brightness adjustment, and random saturation adjustment are implemented. For optimization, we utilize a stochastic gradient descent (SGD) algorithm with a momentum value of 0.9 and a weight decay factor of 0.0004 as the optimizer. The learning rate is adjusted using a polynomial decay strategy, with the polynomial exponent set to 0.9 and the initial learning rate set to 0.009. Considering the characteristics of different datasets, we set specific training epochs and batch sizes: for the NYUDv2 dataset, the training process spans 250 epochs with each batch containing 6 images; for the SUN-RGBD dataset, the training process spans 200 epochs with each batch containing 4 images. The cross-entropy loss function is employed as the network's loss function.

Table 1. Quantitative Comparison on NYUDv2 dataset.

Models	Backbone	PixAcc	mAcc	mIoU
TSNet [33]	ResNet-34	73.50	59.6	46.1
DensnMTL [34]	ResNet-101	-	-	40.84
SCN[35]	ResNet-152	-	-	50.7
TCANet[36]	ResNet-50	71.3	60.3	47.8
ESANet[37]	ResNet-50	74.42	59.27	46.79
CANet [38]	ResNet-50	77.1	64.6	50.9
Link-RGBD [39]	ResNet-50	76.8	59.6	49.5
AESeg [40]	ResNet-50	77.0	-	50.7
Z-ACN [41]	ResNet-50	75.88	63.55	50.05
FCINet [42]	ResNet-50	75.9	63.2	51.7
Our(CFANet)	ResNet-50	78.47	65.31	53.86

4.3. Comparison with SOTA Methods

Table 1 summarizes the experimental results of recent semantic segmentation studies on the NYUDv2 dataset, including several models with advanced performance in recent years. We also evaluated our proposed CFANet on the same dataset and compared its performance with other state-of-the-art (SOTA) models. The experimental data clearly demonstrates that CFANet achieves outstanding results in the crucial performance metric mIoU, surpassing the latest SOTA model FCINet [42] by approximately 2.16%. This unequivocally indicates the significant effectiveness of the novel modules introduced in CFANet. It is noteworthy that, compared to SCN [35] with ResNet-152 as the backbone network, CFANet exhibits superior performance in mIoU, suggesting that even in a relatively simple network structure, CFANet can enhance the accuracy of semantic segmentation. Compared to TSTNet [33] with ResNet-34 as the backbone network, CFANet achieves improvements of 7.76%, 5.71%, and 4.97% in mIoU, mACC, and PixACC, respectively, highlighting the effectiveness of applying residual connection strategies in multiple modules of CFANet.

On the larger-scale SUN-RGBD dataset, we validated the effectiveness of CFANet, and the relevant experimental data can be found in Table 2. Compared to other models in the table, CFANet exhibits excellent performance in mIoU, mACC, and PixACC, showcasing our model's ability to maintain high segmentation accuracy when handling large-scale datasets. Specifically, compared to FCINet with ResNet-50 as the backbone net-work, CFANet achieves a 2.35% improvement in mIoU. Notably, in comparison to models with ResNet-101 as the backbone network in Table 2, our network, despite employing a relatively shallower feature extraction net-work, still achieves superior segmentation performance. This underscores the remarkable capabilities of CFANet in handling large-scale and rich datasets, providing robust support for its widespread applicability in real-world scenarios.

 doi:10.20944/preprints202504.1682.v1

Table 2. Quantitative Comparison on SUN-RGBD dataset.

Models	Backbone	PixAcc	mAcc	mIoU
CANet [38]	ResNet-101	72.5	60.5	49.3
PDCNet [43]	ResNet101	72.4	-	49.2
SGNet[2]	ResNet-101	71.0	-	47.5
CGBNet[44]	ResNet-101	72.3	-	48.2
CMX-B2 [45]	MiT-B2	72.8	-	49.7
ESANet [46]	ResNet-34	-	-	48.2
Link-RGBD [39]	ResNet-50	73.1	53.5	48.4
ACNet [47]	ResNet-50	-	-	48.1
ESANet [46]	ResNet-50	-	-	48.3
FCINet[42]	ResNet-50	72.6	60.9	49.5
Our（CFANet）	ResNet-50	**83.62**	**64.53**	**51.85**

4.4. Visualization Results

To visually showcase the significant advancements of CFANet in semantic segmentation tasks, we provide visualization results on the NYUDv2 dataset, as depicted in Figure 8. A comparison is made with two state-of-the-art network models.

In the first row, noise is present in the feature extraction of "mirror" predictions from both RGB and depth im-ages. CFANet, by integrating contextual information, accurately achieves the segmentation of the "mirror." In the second row, compared to AESeg utilizing asymmetric convolution, CFANet more accurately segments the object within the red dashed box, highlighting the effectiveness of CFANet's combination of asymmetric convolution and dilated convolution. In the third row, FCINet attempts to enhance the network's segmentation accuracy for objects of different scales using spatial pyramid pooling. However, it performs poorly in segmenting large-scale objects (the object within the red dashed box). In contrast, CFANet achieves relatively advanced results by extracting significant single-modal features from the channel and spatial dimensions of the depth map and interacting with the feature map of the RGB image. In the fourth row, the object within the red dashed box has a similar color to the bed. Both FCINet and AESeg fail to segment this object effectively, while CFANet's segmentation result is superior. This is attributed to the appropriate feature extraction modules applied to RGB images and depth maps. These examples highlight CFANet's superiority in semantic segmentation tasks, emphasizing its effectiveness in fusing multimodal information and segmenting objects of different scales.

 doi:10.20944/preprints202504.1682.v1

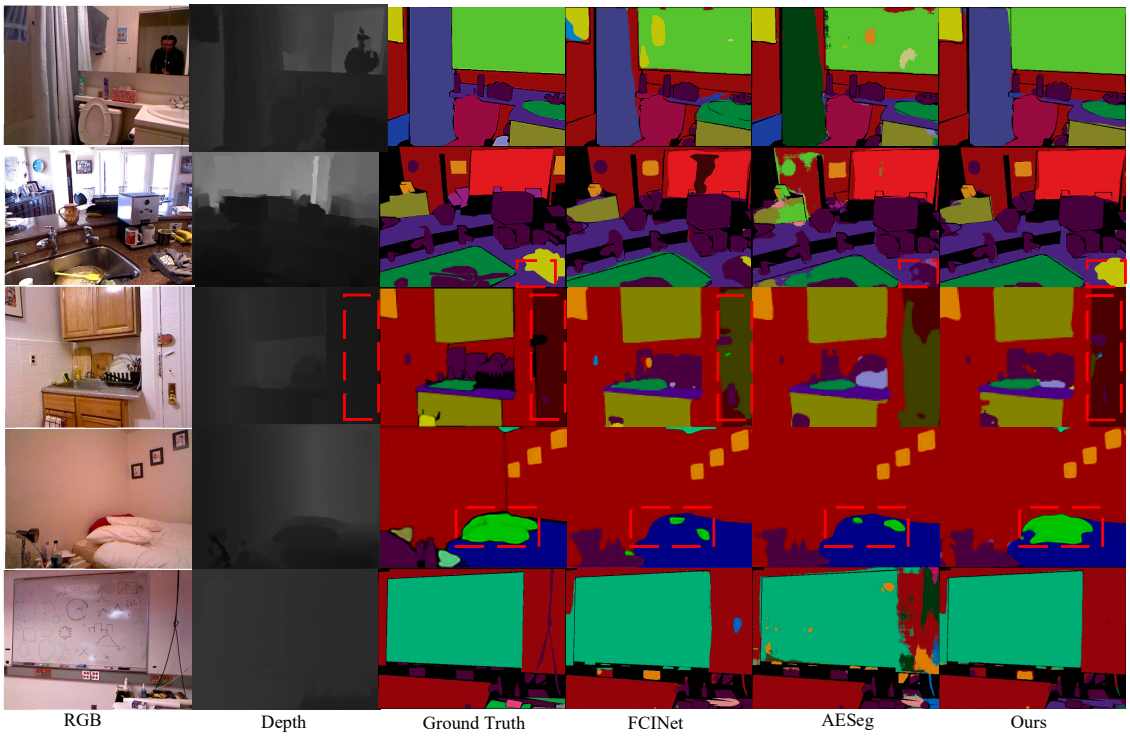


Figure 8. Visualization results on NYUDv2

4.5. Ablation Studies

The CFANet proposed by us comprises several modules, including BFEM, DFEM, AFCFM, SCM and FBH. To validate the effectiveness of each module, we conducted extensive ablation experiments on NYUDv2, and the experimental results are presented in Table 3.

4.5.1. Validate BFEM and DFEM

In this section, we designed four different variants of feature interaction, as illustrated in Figure 9, and obtained experimental results for each variant on the NYUDv2 dataset. The results are recorded in Table 3. By comparing the experimental outcomes of each variant, we substantiate the effectiveness of our research approach.

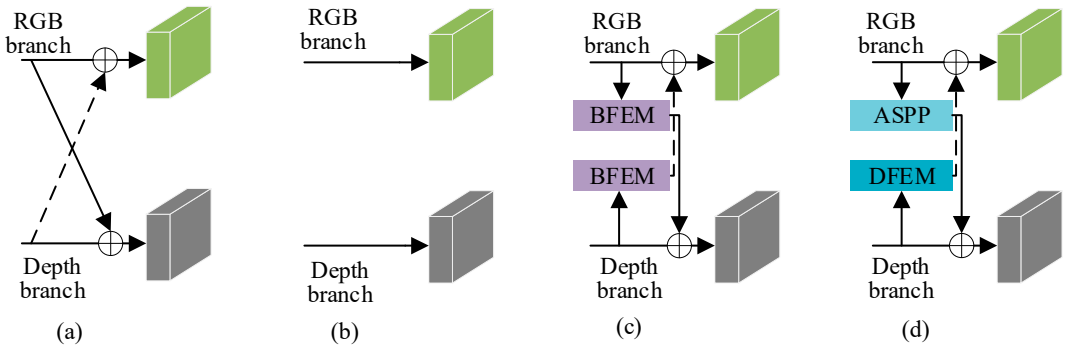


Figure 9. Variants of feature interact.(a) Pixel-wise addition for RGB and depth feature interaction; (b) The feature interaction between the RGB branch and the depth branch is canceled; (c) Using BFEM instead of DFEM to achieve feature re-extraction of the depth branch; (d) Using ASPP instead of BFEM to achieve feature re-extraction of the depth branch.

As illustrated in Figure 9(a), we removed the BFEM and DFEM from CFANet (ResNet-50), and the experimental results are presented in the second row of Table 3, showing a decrease in mIoU by 5.55% compared to CFANet(ResNet-50). In Figure 9(b), we eliminated the feature interaction between the RGB branch and the Depth branch, resulting in the third row of Table 3, with a decrease in mIoU

 doi:10.20944/preprints202504.1682.v1

by 3.88% compared to CFANet (ResNet-50). This reduction is less significant than the strategy shown in Figure 9(a), indicating the existence of a semantic gap between the RGB feature map and the Depth feature map, which leads to poor performance when added directly. Therefore, it is necessary to design appropriate modules to extract features from each modality, as proposed in this paper.

Table 3. Ablation study for various components of CFANet is conducted on the NYUDv2 dataset.

Models	PixAcc	mAcc	mIoU
Without DFEM and BFEM	73.26	62.35	48.31
No interaction between the RGB and depth data	74.16	63.43	49.98
Replace DFEM with BFEM	78.31	64.33	51.81
Replace BFEM with ASPP	78.31	64.29	51.51
Replace DFEM with ASPP	78.28	64.25	51.49
Without AFEM	74.56	63.51	50.23
Without SCM	78.02	64.12	50.75
Without FRH	77.83	64.56	50.94
CFANet(ResNet-50)	78.47	65.31	53.86

As illustrated in Figure 9(c), we replaced the DFEM of the depth branch with BFEM, and the experimental results are presented in the fourth row of Table 3, showing a decrease in mIoU by 2.05% compared to CFANet (ResNet-50). This indicates that the DFEM designed in this paper exhibits certain specificity in extracting features from depth maps. Concurrently, the experimental results also validate the effectiveness of employing appropriate feature extraction modules for the RGB and depth map branches, respectively.

In Figure 9(d), we replaced BFEM of the RGB branch with ASPP, resulting in the fifth row of Table 3, with a de-crease in mIoU by 2.35% compared to CFANet (ResNet-50). BFEM is an improvement upon ASPP, indicating the effectiveness of our enhancement strategy.

Similar to Figure 9(d), we replaced DFEM of the Depth branch with ASPP, resulting in the sixth row of Table 3, with a decrease in mIoU by 2.37% compared to CFANet (ResNet-50). This confirms the effectiveness of DFEM.

4.5.2. Validate AFEM, SCM, and FRH

As shown in the seventh, eighth, and ninth rows of Table 3, we successively replaced AFEM with element-wise summation, SCM with element-wise summation, and removed FRH. The mIoU decreased by 3.63%, 3.11%, and 2.92% compared to CFANet (ResNet-50), respectively. This indicates that AFEM, SCM, and FRH in CFANet (ResNet-50) are effective.

4.5.3. Validate Backbone Network

To evaluate the impact of different backbone networks on the performance of CFANet, this paper retains all modules of CFANet and only replaces the backbone network. As shown in Table 4, this paper selects VGG16, ResNet-18, ResNet-34, and ResNet-50 as alternatives for the backbone network. Experimental results indicate that CFANet with ResNet-50 as the backbone network significantly outperforms the other three variants in terms of mIoU, validating the rationality of using ResNet-50 as the backbone network for CFANet. In contrast, CFANet with ResNet-18 as the backbone network shows a decline in mIoU, which may be due to the limited representation ability of shallower networks for key features, indicating that it is not advisable to excessively reduce net-work depth to minimize model parameters. Similarly, CFANet with ResNet-101 as the backbone network also shows a slight decline in mIoU, which may be due to the loss of crucial feature information caused by blindly increasing the number of network layers. Overall, the experiments demonstrate that ResNet-50 is the most suitable backbone network for CFANet.

 doi:10.20944/preprints202504.1682.v1

Table 4. Experimental results of CFANet based on different backbones.

Backbone	PixAcc	mAcc	mIoU
VGG16	75.17	61.79	48.75
ResNet18	76.75	63.58	50.52
ResNet101	78.02	63.66	52.08
ResNet50	78.47	65.31	53.86

5. Conclusion

In contrast to the conventional homogeneous design of feature extraction modules for both RGB and depth maps, this paper constructs separate feature extraction modules tailored to the characteristics of each. BFEM incorporates asymmetric convolution and cross-cross attention on top of dilated convolutions to alleviate grid-ding effects and capture rich contextual information. DFEM extracts significant unimodal features from both channel and spatial dimensions. For different scales of feature maps, we employ sophisticated strategies: SCM blends multi-scale feature maps, and considering that the first layer contains the richest details while the last layer contains the richest semantic information, we introduce FRH to further fuse and refine these feature maps. CFANet showcases outstanding performance on the NYUDv2 and SUN-RGBD datasets. Visualizations on the NYUDv2 dataset underscore CFANet's exceptional robustness in enhancing multi-scale objects, large-sized objects, and overlapping objects with similar colors. Future research directions will focus on reducing network parameters while maintaining segmentation accuracy and exploring the successful application of CFANet in semantic segmentation for road scenes.

Author Contributions: Conceptualization, L.-F.W. and D.W.; methodology, L.-F.W; software, L.-F.W.; validation, L.-F.W., D.W. and C.-A.X.; formal analysis, D.W.; investigation, C.-A.X.; resources, D.W. and C.-A.X.; data curation, D.W.; writing—original draft preparation, D.W.; writing—review and editing, L.-F.W.; visualization, L.-F.W.; supervision, D.W.; project administration, C.-A.X.; funding acquisition, D.W. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by National Natural Science Foundation of China (No.62101314).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The SUN-RGBD dataset used in this study can be obtained from the website https://rgbd.cs.princeton.edu/ , and the NYUv2 dataset can be obtained from the website https://cs.nyu.edu/~fergus/datasets/nyu_depth_v2.html .

Acknowledgments: Thank you to Shanghai University of Engineering Science for your support of this project

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. El Badaoui, R.; Bonmati Coll, E.; Psarrou, A.; Asaturyan, H.A.; Villarini, B. Enhanced CATBraTS for Brain Tumour Semantic Segmentation. *Journal of Imaging* **2025**, *11*, 8.
2. Chen, L.-Z.; Lin, Z.; Wang, Z.; Yang, Y.-L.; Cheng, M.-M. Spatial Information Guided Convolution for Real-Time RGBD Semantic Segmentation. *Ieee Transactions on Image Processing* **2021**, *30*, 2313-2324, doi:10.1109/tip.2021.3049332.
3. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2017; pp. 2881-2890.
4. Li, X.; He, H.; Li, X.; Li, D.; Cheng, G.; Shi, J.; Weng, L.; Tong, Y.; Lin, Z. Pointflow: Flowing semantics through points for aerial image segmentation. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021; pp. 4217-4226.

 doi:10.20944/preprints202504.1682.v1

5. Cao, J.; Leng, H.; Cohen-Or, D.; Lischinski, D.; Chen, Y.; Tu, C.; Li, Y. RGBxD: Learning depth-weighted RGB patches for RGB-D indoor semantic segmentation. *Neurocomputing* **2021**, *462*, 568-580, doi:10.1016/j.neucom.2021.08.009.
6. Yan, X.; Hou, S.; Karim, A.; Jia, W. RAFNet: RGB-D attention feature fusion network for indoor semantic segmentation. *Displays* **2021**, *70*, doi:10.1016/j.displa.2021.102082.
7. Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *Ieee Transactions on Pattern Analysis and Machine Intelligence* **2018**, *40*, 834-848, doi:10.1109/tpami.2017.2699184.
8. Xiao, L.; Wu, B.; Hu, Y. Surface defect detection using image pyramid. *IEEE Sensors Journal* **2020**, *20*, 7181-7188.
9. Dollár, P.; Appel, R.; Belongie, S.; Perona, P. Fast feature pyramids for object detection. *IEEE transactions on pattern analysis and machine intelligence* **2014**, *36*, 1532-1545.
10. Kazerouni, A.; Karimijafarbigloo, S.; Azad, R.; Velichko, Y.; Bagci, U.; Merhof, D. Fusenet: self-supervised dual-path network for medical image segmentation. In Proceedings of the 2024 IEEE International Symposium on Biomedical Imaging (ISBI), 2024; pp. 1-5.
11. Zheng, Z.; Xie, D.; Chen, C.; Zhu, Z. Multi-resolution Cascaded Network with Depth-similar Residual Module for Real-time Semantic Segmentation on RGB-D Images. In Proceedings of the 2020 IEEE International Conference on Networking, Sensing and Control (ICNSC), 30 Oct.-2 Nov. 2020, 2020; pp. 1-6.
12. Zhou, W.; Yang, E.; Lei, J.; Yu, L. FRNet: Feature Reconstruction Network for RGB-D Indoor Scene Parsing. *Ieee Journal of Selected Topics in Signal Processing* **2022**, *16*, 677-687, doi:10.1109/jstsp.2022.3174338.
13. Zhou, W.; Yuan, J.; Lei, J.; Luo, T. TSNet: Three-Stream Self-Attention Network for RGB-D Indoor Semantic Segmentation. *Ieee Intelligent Systems* **2021**, *36*, 73-78, doi:10.1109/mis.2020.2999462.
14. Cao, J.; Leng, H.; Lischinski, D.; Cohen-Or, D.; Tu, C.; Li, Y.; Ieee. ShapeConv: Shape-aware Convolutional Layer for Indoor RGB-D Semantic Segmentation. In Proceedings of the 18th IEEE/CVF International Conference on Computer Vision (ICCV), Electr Network, 2021 Oct 11-17, 2021; pp. 7068-7077.
15. Zhang, G.; Xue, J.H.; Xie, P.; Yang, S.; Wang, G. Non-Local Aggregation for RGB-D Semantic Segmentation. *IEEE Signal Processing Letters* **2021**, *28*, 658-662, doi:10.1109/LSP.2021.3066071.
16. Zhou, W.; Yue, Y.; Fang, M.; Qian, X.; Yang, R.; Yu, L. BCINet: Bilateral cross-modal interaction network for indoor scene understanding in RGB-D images. *Information Fusion* **2023**, *94*, 32-42, doi:10.1016/j.inffus.2023.01.016.
17. Mnih, V.; Heess, N.; Graves, A.; Kavukcuoglu, K. Recurrent models of visual attention. *Advances in neural information processing systems* **2014**, *27*.
18. Jaderberg, M.; Simonyan, K.; Zisserman, A. Spatial transformer networks. *Advances in neural information processing systems* **2015**, *28*.
19. Huang, Z.; Wang, X.; Huang, L.; Huang, C.; Wei, Y.; Liu, W. Ccnet: Criss-cross attention for semantic segmentation. In Proceedings of the Proceedings of the IEEE/CVF international conference on computer vision, 2019; pp. 603-612.
20. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2018; pp. 7132-7141.
21. Qin, Z.; Zhang, P.; Wu, F.; Li, X. Fcanet: Frequency channel attention networks. In Proceedings of the Proceedings of the IEEE/CVF international conference on computer vision, 2021; pp. 783-792.
22. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. ECA-Net: Efficient channel attention for deep convolutional neural networks. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020; pp. 11534-11542.
23. Yang, Z.; Zhu, L.; Wu, Y.; Yang, Y. Gated channel transformation for visual recognition. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020; pp. 11794-11803.
24. Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the Proceedings of the European conference on computer vision (ECCV), 2018; pp. 3-19.
25. Hu, J.; Wang, H.; Wang, J.; Wang, Y.; He, F.; Zhang, J. SA-Net: A scale-attention network for medical image segmentation. *PloS one* **2021**, *16*, e0247388.

 doi:10.20944/preprints202504.1682.v1

26. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual attention network for scene segmentation. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019; pp. 3146-3154.
27. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* **2020**.
28. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the Proceedings of the IEEE/CVF international conference on computer vision, 2021; pp. 10012-10022.
29. Dong, X.; Bao, J.; Chen, D.; Zhang, W.; Yu, N.; Yuan, L.; Chen, D.; Guo, B. Cswin transformer: A general vision transformer backbone with cross-shaped windows. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022; pp. 12124-12134.
30. Ding, X.; Guo, Y.; Ding, G.; Han, J. Acnet: Strengthening the kernel skeletons for powerful cnn via asymmetric convolution blocks. In Proceedings of the Proceedings of the IEEE/CVF international conference on computer vision, 2019; pp. 1911-1920.
31. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Advances in neural information processing systems* **2017**, *30*.
32. Ruan, H.; Tan, Z.; Chen, L.; Wan, W.; Cao, J. Efficient sub-pixel convolutional neural network for terahertz image super-resolution. *Optics letters* **2022**, *47*, 3115-3118.
33. Zhou, W.; Yuan, J.; Lei, J.; Luo, T. TSNet: Three-stream self-attention network for RGB-D indoor semantic segmentation. *IEEE intelligent systems* **2020**, *36*, 73-78.
34. Lopes, I.; Tuan-Hung, V.; de Charette, R.; Ieee. Cross-task Attention Mechanism for Dense Multi-task Learning. In Proceedings of the 23rd IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, 2023 Jan 03-07, 2023; pp. 2328-2337.
35. Lin, D.; Zhang, R.; Ji, Y.; Li, P.; Huang, H. SCN: Switchable context network for semantic segmentation of RGB-D images. *IEEE transactions on cybernetics* **2018**, *50*, 1120-1131.
36. Jia, W.; Yan, X.; Liu, Q.; Zhang, T.; Dong, X. TCANet: three-stream coordinate attention network for RGB-D indoor semantic segmentation. *Complex & Intelligent Systems* **2024**, *10*, 1219-1230.
37. Seichter, D.; Köhler, M.; Lewandowski, B.; Wengefeld, T.; Gross, H.-M. Efficient rgb-d semantic segmentation for indoor scene analysis. In Proceedings of the 2021 IEEE international conference on robotics and automation (ICRA), 2021; pp. 13525-13531.
38. Zhou, H.; Qi, L.; Huang, H.; Yang, X.; Wan, Z.; Wen, X. CANet: Co-attention network for RGB-D semantic segmentation. *Pattern Recognition* **2022**, *124*, doi:10.1016/j.patcog.2021.108468.
39. Wu, P.; Guo, R.; Tong, X.; Su, S.; Zuo, Z.; Sun, B.; Wei, J. Link-RGBD: Cross-Guided Feature Fusion Network for RGBD Semantic Segmentation. *Ieee Sensors Journal* **2022**, *22*, 24161-24175, doi:10.1109/jsen.2022.3218601.
40. Zhou, W.; Xiao, Y.; Qiang, F.; Dong, X.; Xu, C.; Yu, L. AESeg: Affinity-Enhanced Segmenter Using Feature Class Mapping Knowledge Distillation for Efficient RGB-D Semantic Segmentation of Indoor Scenes. *Neural Networks* **2025**, 107438.
41. Wu, Z.; Allibert, G.; Stolz, C.; Ma, C.; Demonceaux, C. Depth-Adapted CNNs for RGB-D Semantic Segmentation. *Arxiv* **2022**, doi:arXiv:2206.03939.
42. Liu, H.; Xie, W.; Wang, S. Feature fusion and context interaction for RGB-D indoor semantic segmentation. *Applied Soft Computing* **2024**, *167*, 112379.
43. Yang, J.; Bai, L.; Sun, Y.; Tian, C.; Mao, M.; Wang, G.J.a.e.-p. Pixel Difference Convolutional Network for RGB-D Semantic Segmentation. **2023**, arXiv:2302.11951, doi:10.48550/arXiv.2302.11951.
44. Ding, H.; Jiang, X.; Shuai, B.; Liu, A.Q.; Wang, G. Semantic Segmentation With Context Encoding and Multi-Path Decoding. *Ieee Transactions on Image Processing* **2020**, *29*, 3520-3533, doi:10.1109/tip.2019.2962685.
45. Zhang, J.; Liu, H.; Yang, K.; Hu, X.; Liu, R.; Stiefelhagen, R.J.a.e.-p. CMX: Cross-Modal Fusion for RGB-X Semantic Segmentation with Transformers. **2022**, arXiv:2203.04838, doi:10.48550/arXiv.2203.04838.

 doi:10.20944/preprints202504.1682.v1

46. Seichter, D.; Koehler, M.; Lewandowski, B.; Wengefeld, T.; Gross, H.-M.; Ieee. Efficient RGB-D Semantic Segmentation for Indoor Scene Analysis. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), Xian, PEOPLES R CHINA, 2021 May 30-Jun 05, 2021; pp. 13525-13531.
47. Hu, X.; Yang, K.; Fei, L.; Wang, K.; Ieee. ACNET: ATTENTION BASED NETWORK TO EXPLOIT COMPLEMENTARY FEATURES FOR RGBD SEMANTIC SEGMENTATION. In Proceedings of the 26th IEEE International Conference on Image Processing (ICIP), Taipei, TAIWAN, 2019 Sep 22-25, 2019; pp. 1440-1444.