sensors

MDPI

Article

CMANet: Cross-Modality Attention Network for Indoor-Scene Semantic Segmentation

Longze Zhu [1], Zhizhong Kang [1,*], Mei Zhou [2], Xi Yang [3], Zhen Wang [1], Zhen Cao [1] and Chenming Ye [1]

[1] School of Land Science and Technology, China University of Geosciences, Beijing 100083, China
[2] Key Laboratory of Quantitative Remote Sensing Information Technology, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China
[3] College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou 310023, China
* Correspondence: zzkang@cugb.edu.cn; Tel.: +86-136-7128-1065

Abstract: Indoor-scene semantic segmentation is of great significance to indoor navigation, high-precision map creation, route planning, etc. However, incorporating RGB and HHA images for indoor-scene semantic segmentation is a promising yet challenging task, due to the diversity of textures and structures and the disparity of multi-modality in physical significance. In this paper, we propose a Cross-Modality Attention Network (CMANet) that facilitates the extraction of both RGB and HHA features and enhances the cross-modality feature integration. CMANet is constructed under the encoder–decoder architecture. The encoder consists of two parallel branches that successively extract the latent modality features from RGB and HHA images, respectively. Particularly, a novel self-attention mechanism-based Cross-Modality Refine Gate (CMRG) is presented, which bridges the two branches. More importantly, the CMRG achieves cross-modality feature fusion and produces certain refined aggregated features; it serves as the most crucial part of CMANet. The decoder is a multi-stage up-sampled backbone that is composed of different residual blocks at each up-sampling stage. Furthermore, bi-directional multi-step propagation and pyramid supervision are applied to assist the leaning process. To evaluate the effectiveness and efficiency of the proposed method, extensive experiments are conducted on NYUDv2 and SUN RGB-D datasets. Experimental results demonstrate that our method outperforms the existing ones for indoor semantic-segmentation tasks.

Keywords: semantic segmentation; indoor scene; HHA data; cross-modality aggregation; attention mechanism

Citation: Zhu, L.; Kang, Z.; Zhou, M.; Yang, X.; Wang, Z.; Cao, Z.; Ye, C. CMANet: Cross-Modality Attention Network for Indoor-Scene Semantic Segmentation. *Sensors* **2022**, *22*, 8520. https://doi.org/10.3390/s22218520

Academic Editor: Jiayi Ma

Received: 14 September 2022
Accepted: 2 November 2022
Published: 5 November 2022

1. Introduction

Semantic segmentation is one of the essential techniques in scene understanding technologies. It aims to categorize each pixel and assists in the identification and segmentation of scene elements. Initially, the segmentation can be achieved by handcrafted features and machine-learning algorithms [1–3], among which, deep learning is the trend of current research [4–7]. For semantic-segmentation tasks, it is fully recognized that indoor scenes exhibit distinct aspects compared with outdoor scenarios. For instance, services relying on indoor-scene semantic segmentation (indoor navigation, intelligent furniture, etc.) are strongly demanded by individuals. Therefore, facing the aforementioned challenges, some works apply notable features of the images to assist the segmentation, e.g., edges information [8,9]. Additionally, the illumination variations, overlaps among objects, and the imbalanced representations of object categories in indoor scenes always make it impossible to distinguish numerous objects using solely RGB images [10].

Adding the depth information to traditional RGB information with low-cost RGB-D sensors is a conventional way to achieve better performance of indoor-scene semantic segmentation. Nevertheless, the depth images contain only range measurement information, which makes them challenging for feature extraction. As a result, it is natural to employ the three-channel HHA images (three channels represent the horizontal disparity, height

above ground, and the angle of the pixel's local surface normal makes with the inferred gravity direction) [11], which are coded from one-channel depth images. The application of HHA images is demonstrated to be more efficient and robust than one-channel depth images [4,12,13]. The comparison of RGB, depth, and HHA images are shown in Figure 1.

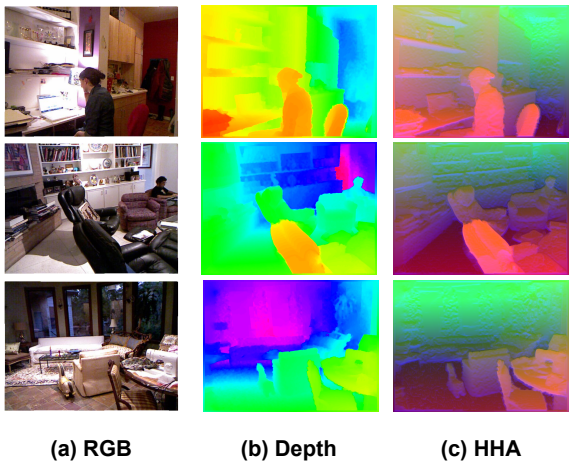

Figure 1. Comparisons: (**a**) RGB images, (**b**) one-channel depth images, and (**c**) three-channel HHA images.

In order to improve feature embedding, it is necessary to consider the essence of data modality, i.e., the quality, attribute, or circumstance of each type. The RGB images describe lightness and saturation, which mainly represent appearance information, whereas the HHA images mainly represent geometric information [14]. As a result, RGB and HHA can complement one another, with HHA discriminating instances and contexts that share similar colors and textures [15], while RGB can assist with indistinguishable structures. As illustrated in Figure 2, the cushion on the sofa has a similar texture to the sofa, so it can be distinguished by the HHA images, and the pictures on the wall have a similar structure as the wall, so they can be distinguished by the RGB images. It is evident that the combination of both modality features from RGB and HHA images can effectively enhance the efficiency of feature embedding.

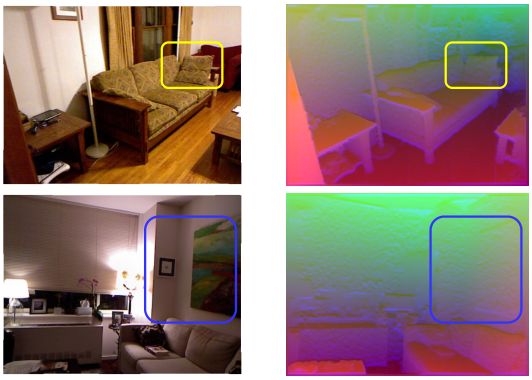

Figure 2. Some challenging samples of semantic-segmentation tasks. The yellow bounding box indicates the complex sample on solely RGB images, whereas the blue indicates the complex sample with solely HHA images.

On the basis of existing deep-learning methods for RGB-D semantic segmentation, two open problems are still widely discussed: how to fuse multi-modality (RGB and depth) features adeptly and how to improve the robustness w.r.t. imperfect data. The first problem is caused by the substantial variations between RGB and depth modalities [12], which lead to inappropriate feature fusion and inferior performance. Some works utilize the depth

image as an extra channel [4,16,17], whereas some works extract features independently and fuse them via CNN-based architecture [14,18,19]. Despite this, these methods can only correlate RGB and depth information to a limited extent. Additionally, measurement noise, view angle, and occlusion boundary may affect the RGB sensors (e.g., resulting in overexposure) and depth sensors (e.g., resulting in data loss) during the data collection period, causing the second problem. Therefore, some works [20,21] apply pre-processing to RGB and depth images for denoising and filling in lost depth. However, pre-processing is not stable enough for feature extraction.

In this paper, we propose a novel Cross-Modality Attention Network (CMANet) for indoor-scene semantic segmentation. CMANet is designed under the encoder–decoder architecture. The encoder aims to build multi-step interaction between two modalities and extracts multi-level features from RGB and HHA images. Meanwhile, the decoder enhances the efficiency of feature representation and restores the feature maps to the corresponding resolution step-by-step. It is worth mentioning that the cross-modality ability ensures that CMANet can utilize the essence of different modalities effectively and fuse the RGB and HHA features adeptly. To be specific, the encoder has two parallel branches to extract RGB and HHA features. The parallel design scheme minimizes the unfavorable effects of mutual influence between RGB and HHA features. More importantly, to achieve appropriate feature fusion and filter out the noise of images, we propose the Cross-Modality Refine Gate (CMRG) based on the attention mechanism, weighting the crucial features in the first stage and aggregating them in the second stage. Moreover, the CMRG utilizes the features from different modalities, instead of relying on one modality, which increases the model robustness and achieves a great performance. Additionally, the output of the CMRG is propagated in a multi-step bi-directional operation into both branches to enhance the encoding of features. The decoder is an up-sampled ResNet, which gradually restores spatial resolution and integrates the encoding stage information. Additionally, we conduct pyramid supervision to improve the final semantic performance.

The main contributions of this paper are as follows:

- We propose a novel RGB-HHA semantic-segmentation network: CMANet. On one hand, CMANet can effectively fuse and extract the cross-modality features; on the other hand, it improves the robustness and efficiency of feature embedding;
- We design the CMRG based on the self-attention mechanism, which not only filters the noise and highlights advantages in the feature maps, but also facilitates to improve the representation and aggregation of cross-modality information;
- We conduct extensive experiments on challenging NYUDv2 and SUN RGB-D datasets, on which CMANet evidences its robustness and effectiveness on indoor-scene semantic segmentation.

The rest of the paper is organized as follows. Section 2 briefly reviews the related works. Section 3 gives a detailed description of CMANet, which includes the network architecture, the processing modules, the training and optimization strategies, etc. Section 4 provides the experiment settings, evaluation methods, and the results, along with comparing the proposed method with existing methods and analyzing the strengths and limitations of our approach. Section 5 closes the paper with a brief conclusion and future research considerations.

2. Related Works

In this section, we briefly review the literature relevant to our work. The attention mechanism part focuses on the utilization of channel attention and spatial attention, and the RGB-D semantic-segmentation part elaborates the existing methods and essential concerns of cross-modality fusion.

2.1. Attention Mechanism

The attention mechanism is derived from the behavior of humans, which is to ignore irrelevant information and pay attention to what is essential. This strategy is commonly

applied in the deep-learning field and performs well on Natural Language Processing (NLP) [22–24] and Computer Vision (CV) [15,25,26] tasks. Considering the attention mechanism in the CV field, we divide it into two categories based on enhancement type: channel attention (what to focus on) and spatial attention (where to focus on).

In deep learning, the feature maps that propagate in different channels usually represent different features [27], such as texture, boundary, and shape. As a result, channel attention assigns different weights to different channels, determining what features are important (what to focus on), and is widely used [28–33]. SENet [31] provides the first step to improving the representation ability of feature maps through channel attention by using the Squeeze-and-Excitation (SE) block. For better modeling capability, GSoP-Net [32] generates attention maps not only by utilizing first-order statistics (i.e., the global pooling descriptor) but by extracting high-order statistics as well. SKNet [33] combines different receptive field features with multiple branches to adapt the weights according to the input feature maps. These methods open up a novel aspect to feature extraction where the valuable information from existing channels is emphasized rather than continuously increasing the number of parameters. Thus, we adopt the channel-attention mechanism in the CMRG to refine the cross-modality feature maps so as to reduce the noise and improve feature representation.

Regarding one feature map, the region attribute can be represented by pixel values in different locations. The spatial attention can adeptly select the interrelated and meaningful regions across the entire map, then reinforce them by allocating higher weights (where to focus on), which has significance for image processing [34–38]. AGNet [36] applies Attention Gates (AGs) for medical image segmentation. The AGs restrain irrelevant regions and highlight significant features via implicit learning. PSANet [37] relates all the positions on the feature map to each other via a self-adaptively learned attention mask, thus alleviating the limitations of convolutional filters with small kernel size. Additionally, transformers have proven to be very effective with NLP and CV tasks [22,39]; therefore, the Vision Transformer (ViT) [38] processes images by cropping each image to 16×16 small samples, which is similar to split a sentence into several words. The spatial attention mainly concerns the crucial regions in one feature map, which means it can enhance the relationship among pixels. Inspired by these works, we consider the fact that RGB and HHA have different geometric properties and propose a cross-modality spatial-attention mechanism that can enhance region characteristics.

In order to better refine the feature maps, CBAM [40] combines the channel and spatial-attention modules, which will be discussed later. Motivated by this, we employ the sequential deployment of the channel and spatial modules for cross-modality features. In this way, the acquisition of what is essential comes not only from latent learning, but also from different physical characteristics from different modalities.

2.2. RGB-D Semantic Segmentation

For RGB-D semantic segmentation, not only the category labels are assigned to each pixel; the performance is also enhanced via the cross-modality feature fusion. Based on the existing successful deep-learning RGB semantic-segmentation structures [4–7], several works on RGB-D semantic segmentation are presented and demonstrate their reliability and practicality.

Since the RGB and depth modalities must be fully utilized in RGB-D semantic segmentation, the cross-modality fusion is crucial. The fusion can be achieved via element-wise summation, concatenation, or a combination of both, and adapted by latent learning [4,13,14,18,19,41–43]. For learning common and specific parts from cross-modality features, ref. [14] designs the transformation network between the encoder and decoder, and extracts corresponding features via Multiple Kernel Maximum Mean Discrepancy (MK-MMD). RedNet [18] integrates ResNet [44] and encoder–decoder architecture. In addition, it adds skip connections to optimize decoding and applies pyramid supervision to avoid overfitting. Learning from the structure of RefineNet [45], RDF [13] fuses cross-modality features by applying a Multi-Modal feature Fusion (MMF) network, which

is composed of several residual convolutional layers and element-wise summation. To avoid noisy and chaotic information affecting the effectiveness of the network, RAFNet [43] employs a three-stem branch encoder to process RGB, depth, and fusion features, respectively. Meanwhile, RAFNet utilizes the channel-attention model for refinement. The RGB-D semantic-segmentation methods mostly inherit the former deep-learning methods used for image processing, some of which insert the attention mechanism to enhance the features adeptly by filtering noisy and chaotic information. However, the relationship between RGB and depth features has received little interest.

Thus, we propose several processing modules for cross-modality features based on residual convolution and attention mechanism. In this structure, the residual convolution contributes to the latent learning and adjusts adaptively for feature maps, while the attention mechanism contributes to information refinement.

3. The Cross-Modality Attention Network

An effective cross-modality network aims to attenuate the image noise and combine the benefits of both RGB and HHA features. Regarding this, a novel attention-based mechanism is introduced in our proposed model (CMANet), and it results in the so-called CMRG design. Additionally, bi-directional multi-step propagation and pyramid-supervision training strategies facilitate the performance. In this section, we will present detailed descriptions of the proposed method in terms of the overall framework, the structure of the Cross-Modality Refine Gate (CMRG), the processing modules, the configuration of the encoder and decoder, and the pyramid-supervision strategy.

3.1. Network Architecture

In this subsection, we detail the structure of our CMANet, which includes the overall framework, the processing modules for cross-modality features, and the configuration of the encoder and the decoder.

3.1.1. Overall Framework

Influenced by SegNet [5], our proposed structure is based on encoder–decoder architecture. In the encoder part, latent modality features are extracted from RGB and HHA images, which serve as the input of the decoder. Following that, the decoder gradually reconstructs the high-dimensional features to the original spatial resolution and integrates the encoding stage information through skip connections to produce the results of semantic segmentation. In this structure, the encoder extracts the high-level features while the decoder restores the spatial information, which alleviates the problem of chaos in the semantic assignment in pixel-wise classification.

The architecture of CMANet is presented in Figure 3. The encoder has two CNN branches w.r.t. RGB and HHA. Each branch successively extracts latent modality features from RGB and HHA images. Here, ResNet [44] serves as the backbone for both branches.

In order to enhance the extraction and fusion of cross-modality characteristics, we present the Cross-Modality Refine Gate (CMRG), which is designed based on the self-attention mechanism. The CMRG module receives pairs of encoding feature maps from the RGB and HHA branches (e.g., the outputs of RGB-Layer1 and HHA-Layer1) and produces aggregated features. Regarding the cross-modality fusion, there are several CMRG modules, which correspond to distinct encoding stages. As a result, we can extract more valuable features with the availability of an encoder capable of fusing and boosting features via the separate implementation of encoding and fusion with CMRGs.

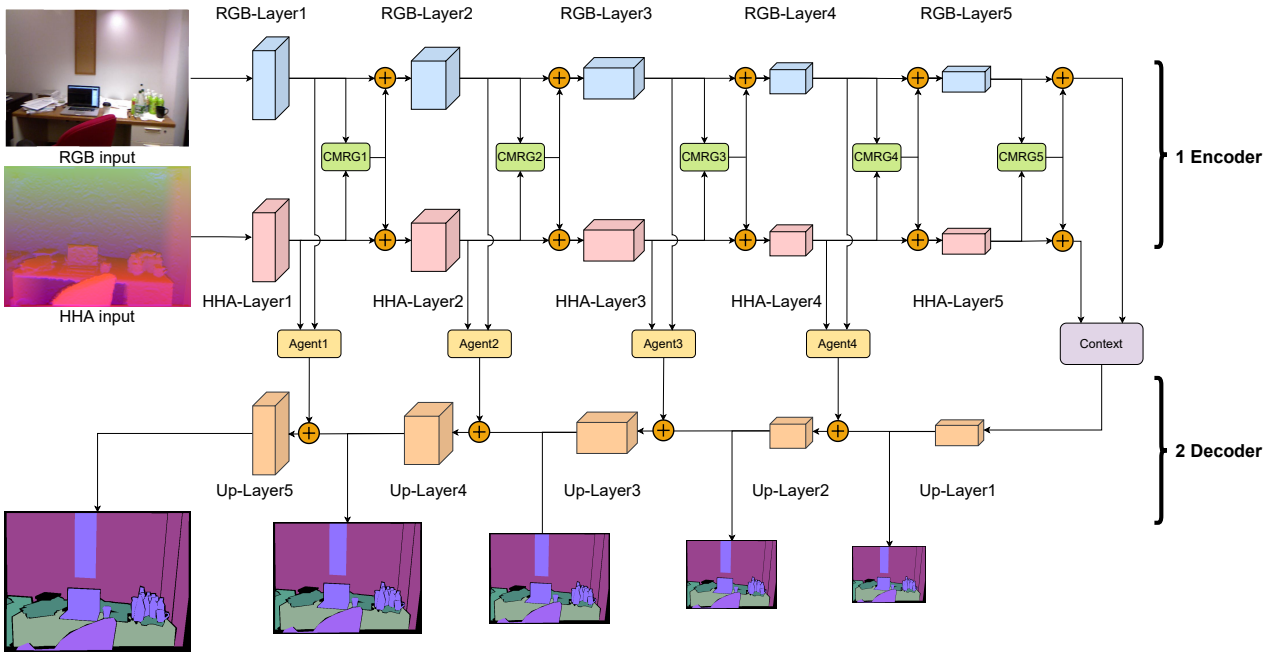

Figure 3. Illustration of the framework of CMANet. CMANet has the encoder–decoder architecture: (**1**) The encoder extracts RGB and HHA features with ResNet [44] backbone. (**2**) The decoder is an up-sampled backbone composed of several standard residual blocks.

After encoding, CMRG5 refines the outputs of RGB-Layer5 and HHA-Layer5 to produce the final pair of feature maps. Nonetheless, the final pairings are the outputs of the encoder, which have different representational capabilities because they are derived from separate modalities and, hence, they cannot be joined by element-wise addition. Consequently, we employ the Context module to aggregate and refine the pairings of high-level feature maps from two modalities.

The decoder component is an up-sampled ResNet backbone consisting of five residual blocks, each of which comprises several Up-sampled Residual Units (URUs). In this structure, the decoder recovers the feature maps to the original spatial resolution stage-by-stage via transposed convolution and combines the features from each encoding stage via Agent modules as skip connections.

The low-level features in the decoder have a better resolution with more position and detail information, but the high-level features contain rich semantic and category information. To improve the use of multi-level features and alleviate the gradient vanishing problem, we generate multi-scale semantic maps from five stages of decoding features for pyramid supervision.

3.1.2. Processing Modules

Here, we outline the processing modules that aid in the propagation of features.

Residual Units The residual learning can effectively prevent the degradation of the model and resolve the gradient vanishing issue during back-propagation. The structure of residual units is shown in Figure 4. The Residual Convolutional Unit (RCU) and the Chained Residual Pooling (CRP) are utilized in the Agent and Context modules, which are the sub-components of RefineNet [45], whereas the Downsample Residual Unit (DRU) [18] and Upsample Residual Unit (URU) [44] are applied for the encoder and decoder, respectively. It is worth mentioning that, while the displayed Chained Residual Pooling (CRP) has two blocks in Figure 4b, we only utilize one block in our subsequent modules since one is sufficient for refinement.

Agent Module The skip connections between the encoder and decoder are utilized to replenish the detail loss caused by downsampling. Hence, the Agent modules provide certain intermediate addition of the multi-stage feature maps from the encoder to the

corresponding decoder layers. The structure of the Agent module is illustrated in Figure 5a. After receiving two feature maps from the RGB and HHA branches, the Agent module first utilizes a 1×1 convolution layer to mitigate the explosion of parameters by reducing the dimension number. Then, each feature map goes through the RCU to adapt the element-wise sum fusion. The combined features are fed to a one-block CRP, where a pooling algorithm spreads large activation values while an additional convolution layer is added to learn the significance of the pooled features. Finally, before passing to the decoder, a Convolutional Block Attention Module (CBAM) is employed to filter and enhance the features. Note that, in order to improve the propagation, we couple the CBAM and residual learning via a shortcut connection.

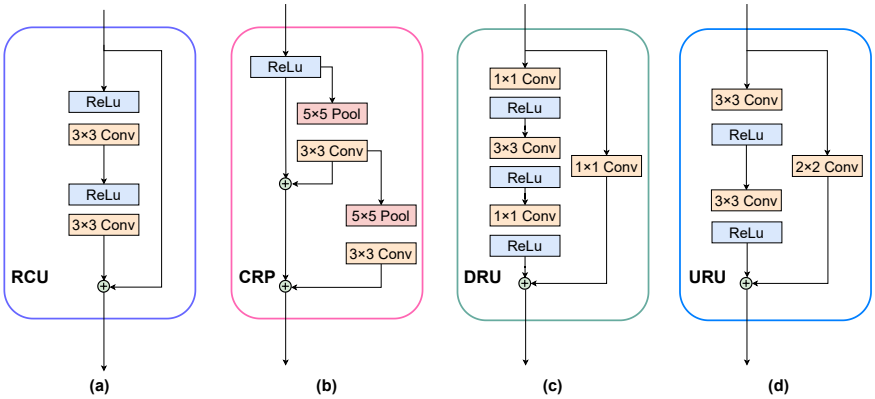

Figure 4. Structure of residual units: (**a**) Residual Convolutional Unit (RCU): a standard residual convolutional unit, which is an adaptive convolution that has two standard 3×3 layers with shortcut connection; (**b**) Chained Residual Pooling (CRP): a chain of pooling blocks (two blocks) with shortcut connection, each of which contains of a 5×5 max-pooling layer and a 3×3 convolution layer; (**c**) Downsample Residual Unit (DRU): a downsample residual unit in the (ResNet-50) backbone; (**d**) Upsample Residual Unit (URU): an upsample residual unit that we propose in the decoder.

Context Module The Context module is applied to fuse the final outputs of the two CNN branches (RGB and HHA). As illustrated in Figure 5b, the Context module has similar components as the Agent module, but has additional RCUs and a 3×3 convolution layer. The first 1×1 convolution layer reduces the dimension from 2048 to 512. Then, two feature maps are fused by element-wise summation and finally output to the decoder after refinement.

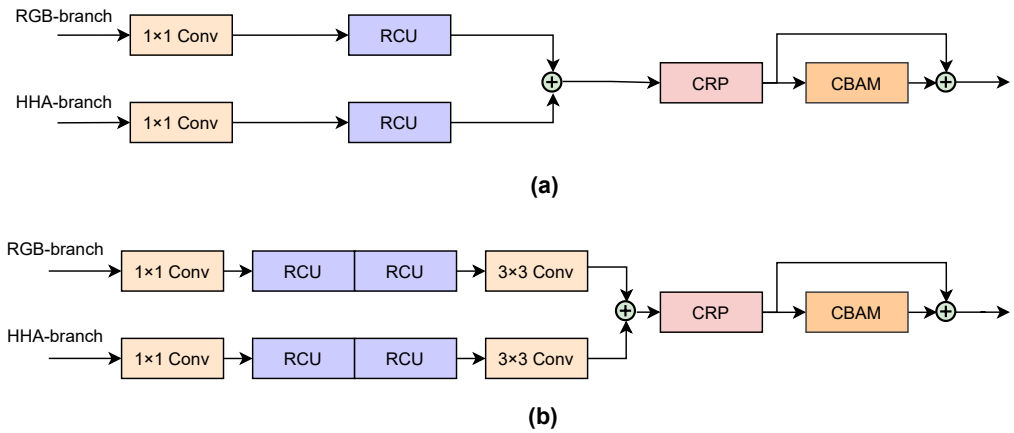

Figure 5. Structures of processing modules: (**a**) the structure of the Agent module; (**b**) the structure of the Context module.

3.1.3. Encoder and Decoder Configuration

The encoder and decoder have different types and numbers of the residual unit. The encoder with the backbone of ResNet-50 utilizes the Downsample Residual Unit (DRU) as illustrated in Figure 4c, with the 1×1 convolution layer with a stride of 2. Regarding the decoder, the residual units are applied to upsample the feature maps, as illustrated in Figure 4d, where the 2×2 convolution layer and the second 3×3 convolution layer both have a stride of $1/2$. The encoder and decoder configuration is shown in Table 1, $Input$ denotes the number of input feature channels, $Output$ denotes the number of output feature channels, and $Units$ denotes the number of residual units in this layer.

Table 1. Encoder (ResNet-50) and decoder configuration.

Block	Encoder			Block	Decoder		
	Input	*Output*	*Units*		*Input*	*Output*	*Units*
Layer1	3	64	-	Trans1	512	256	6
Layer2	64	256	3	Trans2	256	128	4
Layer3	256	512	4	Trans3	128	64	3
Layer4	512	1024	6	Trans4	64	64	3
Layer5	1024	2048	3	Trans5	64	64	3

3.2. Cross-Modality Refine Gate

For RGB and HHA data, the former mainly record appearance information (e.g., color, texture) that can emphasize the visual boundary, whereas the latter primarily capture shape information (e.g., structure, spatial) that can highlight the geometric boundary. Thus, it is challenging to fully utilize RGB and HHA images via fusion and enhancement of cross-modality features. We propose the Cross-Modality Refine Gate (CMRG) based on the attention mechanism to aggregate features from multiple modalities.

3.2.1. Convolutional Block Attention Module

As discussed in Section 2, the attention mechanism has been extensively used in the CV field for determining where the focus should be placed and for deciding what is valuable. In particular, the Convolutional Block Attention Module (CBAM) combines channel and spatial-attention mechanisms during propagation, an achieves outstanding performance in feature extraction [40]. For further modification of the CBAM, we discuss its overall structure and sub-modules details first.

The structures of the CBAM and its sub-modules are shown in Figure 6. The CBAM refines the input feature maps by sequentially applying one channel-attention module and one spatial-attention module, as illustrated in Figure 6a. Given input feature maps $\boldsymbol{\mathcal{F}}$, the CBAM first infers a 1D channel-attention map $\boldsymbol{\mathcal{M}}_c$ using the channel-attention module and refines the feature maps via channel-wise multiplication; then, it infers a 2D spatial-attention map $\boldsymbol{\mathcal{M}}_s$ using the spatial-attention module and performs spatial-wise multiplication on the channel-refined feature maps to yield the output. The CBAM process can be formulated as follows:

$$\boldsymbol{\mathcal{F}}' = \boldsymbol{\mathcal{M}}_c(\boldsymbol{\mathcal{F}}) \otimes \boldsymbol{\mathcal{F}}, \tag{1}$$

$$\boldsymbol{\mathcal{F}}_{out} = \boldsymbol{\mathcal{M}}_s(\boldsymbol{\mathcal{F}}') \otimes \boldsymbol{\mathcal{F}}', \tag{2}$$

where $\boldsymbol{\mathcal{F}} \in \mathbb{R}^{C \times H \times W}$ represents the input feature maps, $\boldsymbol{\mathcal{F}}'$ represents the channel-refined feature maps, $\boldsymbol{\mathcal{F}}_{out}$ represents the final output, and $\otimes$ denotes element-wise multiplication. In the multiplication procedure, the attention values are broadcast as follows. Channel-attention values are broadcast along the spatial dimension (channel-wise multiplication), while spatial-attention values are broadcast along the channel dimension (spatial-wise multiplication). Figure 6b,c describe the detailed structure of the channel-attention module and the spatial-attention module.

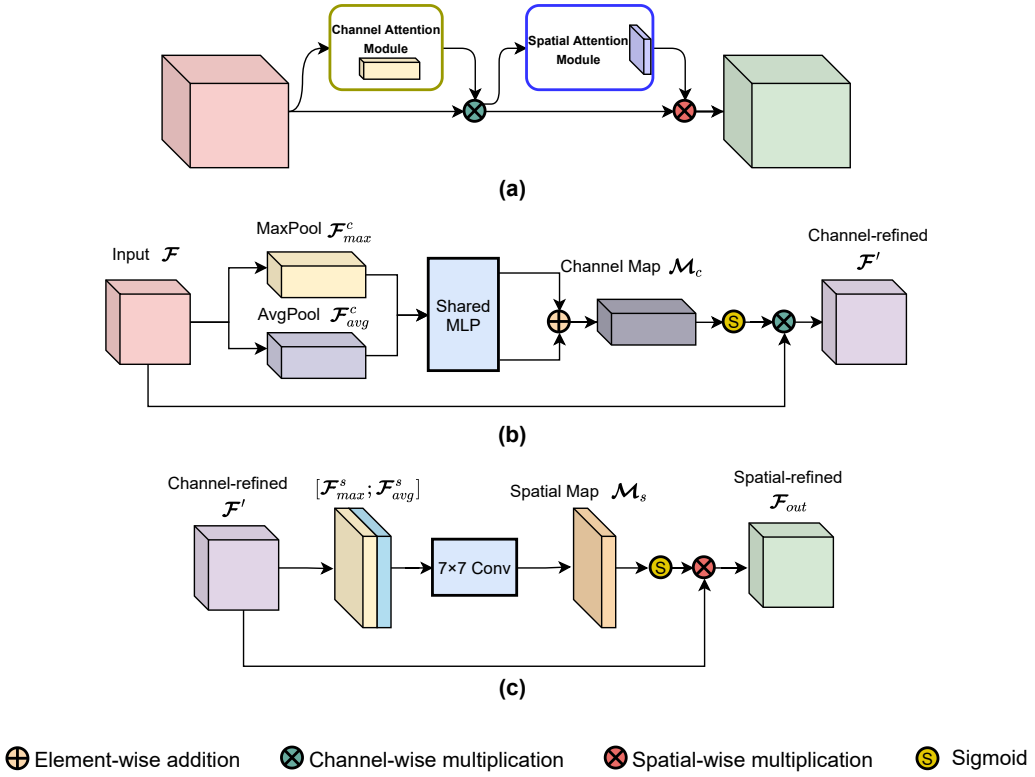

Figure 6. Structures of CBAM and its sub-modules: (**a**) the structure of CBAM, which is comprised of a channel-attention module and a spatial-attention module in sequence; (**b**) the structure of the channel-attention module; (**c**) the structure of the spatial-attention module.

As shown in Figure 6b, the channel-attention module first aggregates the spatial information of input feature $\mathcal{F}$ into two descriptors, average-pooled features $\mathcal{F}^c_{avg}$ and max-pooled features $\mathcal{F}^c_{max}$, via average-pooling and max-pooling operations, respectively. Then, both descriptors are supplied to a shared network containing one hidden layer of multi-layer perception (MLP). A reduction ratio is set to the shared MLP in order to decrease parameter overhead. After propagation in the shared MLP, the descriptors are fused by element-wise summation and modified by a sigmoid function to produce the channel-attention map $\mathcal{M}_c$. At last, the channel-refined features $\mathcal{F}_c$ can be generated by multiplying the attention map and input features. The channel-attention map is computed as follows:

$$\begin{aligned}\mathcal{M}_c(\mathcal{F}) &= \sigma(MLP(AvgPool(\mathcal{F})) + MLP(MaxPool(\mathcal{F}))) \\ &= \sigma(\boldsymbol{W_1}(\boldsymbol{W_0}(\mathcal{F}^c_{avg})) + \boldsymbol{W_1}(\boldsymbol{W_0}(\mathcal{F}^c_{max}))),\end{aligned} \tag{3}$$

where σ denotes the sigmoid function, $\boldsymbol{W_0} \in \mathbb{R}^{C/r \times C}$ and $\boldsymbol{W_1} \in \mathbb{R}^{C \times C/r}$ represent the MLP weights, and r is the reduction ratio. As specified, the MLP weights are shared with both inputs. $\mathcal{M}_c \in \mathbb{R}^{C \times 1 \times 1}$ represents the channel-attention map, $\mathcal{F}^c_{avg} \in \mathbb{R}^{C \times 1 \times 1}$ represents the average-pooled features, and $\mathcal{F}^c_{max} \in \mathbb{R}^{C \times 1 \times 1}$ represents the max-pooled features.

The spatial-attention module is presented in Figure 6c. Similar to the aforementioned channel-attention module, the spatial-attention module aggregates channel information of the refined features to maps $\mathcal{F}^s_{avg}$ and $\mathcal{F}^s_{max}$ first via average-pooling and max-pooling along the channel axis. The two maps are then merged by concatenation. After that, the concatenated features are convolved by a standard convolution layer to generate the spatial-attention map $\mathcal{M}_s$. The final refined features are also produced by multiplication. The spatial-attention map is computed as follows:

$$\begin{aligned}\boldsymbol{\mathcal{M}}_s(\boldsymbol{\mathcal{F}}) &= \sigma(f^{7\times7}([AvgPool(\boldsymbol{\mathcal{F}}); MaxPool(\boldsymbol{\mathcal{F}})]) \\ &= \sigma(f^{7\times7}([\boldsymbol{\mathcal{F}}^s_{avg}; \boldsymbol{\mathcal{F}}^s_{max}])).\end{aligned} \tag{4}$$

where σ denotes the sigmoid function, $f^{7\times7}$ represents the 7×7 convolutional layer, and $[;]$ refers to the concatenation. $\boldsymbol{\mathcal{M}}_s \in \mathbb{R}^{1\times H\times W}$ represents the spatial-attention map, $\boldsymbol{\mathcal{F}}^s_{avg} \in \mathbb{R}^{1\times H\times W}$, and $\boldsymbol{\mathcal{F}}^s_{max} \in \mathbb{R}^{1\times H\times W}$ represent the pooled maps.

3.2.2. Structure of Cross-Modality Refine Gate

Although CBAM exhibits good performance w.r.t. the extraction and representation of features through the channel- and spatial-attention mechanisms, it is inadequate for the cross-modality features. Inspired by CBAM, where the channel- and spatial-attention modules are sequentially employed (and the attention mechanisms are applied with each modality separately), here, we design a unique attention module, the Cross-Modality Refine Gate (CMRG), which is designed to deal with cross-modality features. The structure of the CMRG is illustrated in Figure 7. Differing from the CBAM, the CMRG takes the multi-modality features as the input, instead of receiving only one feature maps. The self-attention mechanism in the CMRG utilizes the cross-modality information to produce the attention maps, which only rely on one modality in CBAM.

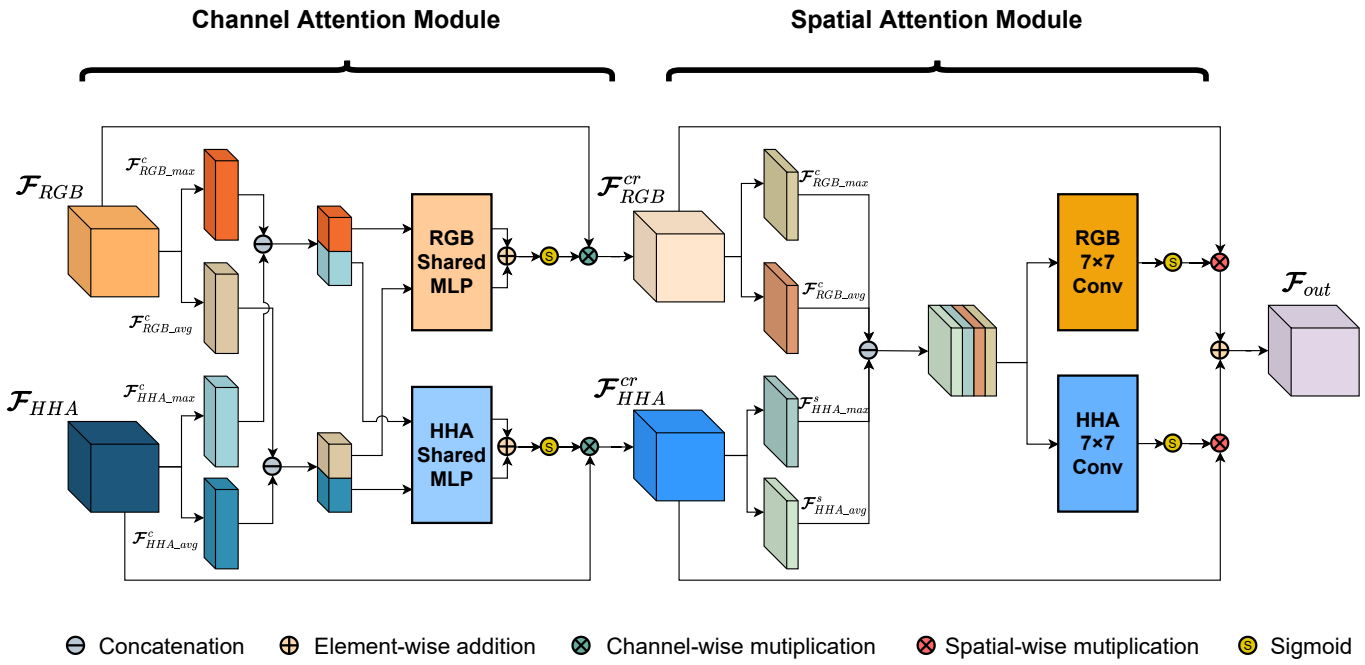

Figure 7. Structure of the Cross-Modality Refine Gate (CMRG).

As shown in Figure 7, the CMRG consists of two parts: the channel-attention module, and the spatial-attention module. The input of the CMRG is a pair of feature maps $\boldsymbol{\mathcal{F}}_{RGB}$ and $\boldsymbol{\mathcal{F}}_{HHA}$, which are derived from the RGB and HHA branches, respectively. Firstly, the CMRG utilizes the channel-attention module to infer two 1D channel-attention maps—$\boldsymbol{\mathcal{M}}^c_{RGB}$ and $\boldsymbol{\mathcal{M}}^c_{HHA}$—to refine $\boldsymbol{\mathcal{F}}_{RGB}$ and $\boldsymbol{\mathcal{F}}_{HHA}$ via channel-wise multiplication. By the sharing of descriptors from each modality and the inference of exclusive attention maps for each modality, the cross-modality channel-attention operation primarily filters out noisy and chaotic information and improves the effective characteristics of the original feature maps from the channel aspect. Then, the CMRG infers two 2D spatial-attention maps—$\boldsymbol{\mathcal{M}}^s_{RGB}$ and $\boldsymbol{\mathcal{M}}^s_{HHA}$—to enhance the channel-refined feature maps $\boldsymbol{\mathcal{F}}^{cr}_{RGB}$ and $\boldsymbol{\mathcal{F}}^{cr}_{HHA}$ via spatial-wise multiplication. The cross-modality spatial attention primarily

strengthens the association among pixels in the feature maps in order to put more attention on the areas with similar characteristics, even though some of them are far from each other. Finally, the output $\mathcal{F}_{out}$ is generated by adding the spatial-refined feature maps together. This process can be formulated as follows:

$$\begin{aligned} \mathcal{F}^{cr}_{RGB} &= \mathcal{M}^{c}_{RGB}(\mathcal{F}_{RGB}, \mathcal{F}_{HHA}) \otimes \mathcal{F}_{RGB}, \\ \mathcal{F}^{cr}_{HHA} &= \mathcal{M}^{c}_{HHA}(\mathcal{F}_{RGB}, \mathcal{F}_{HHA}) \otimes \mathcal{F}_{HHA}, \end{aligned} \tag{5}$$

$$\begin{aligned} \mathcal{F}^{sr}_{RGB} &= \mathcal{M}^{s}_{RGB}(\mathcal{F}^{cr}_{RGB}, \mathcal{F}^{cr}_{HHA}) \otimes \mathcal{F}^{cr}_{RGB}, \\ \mathcal{F}^{sr}_{HHA} &= \mathcal{M}^{s}_{HHA}(\mathcal{F}^{cr}_{RGB}, \mathcal{F}^{cr}_{HHA}) \otimes \mathcal{F}^{cr}_{HHA}, \end{aligned} \tag{6}$$

$$\mathcal{F}_{out} = \mathcal{F}^{sr}_{RGB} + \mathcal{F}^{sr}_{HHA}, \tag{7}$$

where $\mathcal{F}_{RGB} \in \mathbb{R}^{C\times H\times W}$ and $\mathcal{F}_{HHA} \in \mathbb{R}^{C\times H\times W}$ represent the input feature maps, $\mathcal{M}^{c}_{RGB} \in \mathbb{R}^{C\times 1\times 1}$ and $\mathcal{M}^{c}_{HHA} \in \mathbb{R}^{C\times 1\times 1}$ represent the channel-attention maps, $\mathcal{M}^{s}_{RGB} \in \mathbb{R}^{1\times H\times W}$ and $\mathcal{M}^{s}_{HHA} \in \mathbb{R}^{1\times H\times W}$ represent the spatial-attention maps, $\mathcal{F}^{cr}_{RGB} \in \mathbb{R}^{C\times H\times W}$ and $\mathcal{F}^{cr}_{HHA} \in \mathbb{R}^{C\times H\times W}$ represent the channel-refined feature maps, $\mathcal{F}^{sr}_{RGB}$ and $\mathcal{F}^{sr}_{HHA}$ represent the spatial-refined feature maps, and $\otimes$ denotes element-wise multiplication. In the multiplication procedure, the attention values are broadcast the same as in the implementation in CBAM.

As shown in Figure 7, the channel-attention module first aggregates each feature maps into two 1D descriptors, totaling four via average-pooling and max-pooling, among which $\mathcal{F}^{c}_{RGB\_avg}$ and $\mathcal{F}^{c}_{RGB\_max}$ are generated from RGB feature maps, whereas $\mathcal{F}^{c}_{HHA\_avg}$ and $\mathcal{F}^{c}_{HHA\_max}$ are generated from HHA feature maps. Then, the descriptors from various modalities are concatenated to produce two cross-modality channel descriptors: $\mathcal{F}^{c}_{avg}$ and $\mathcal{F}^{c}_{max}$. Both cross-modality descriptors are fed into two independent MLPs with one hidden layer: MLP_{RGB} and MLP_{HHA}. After the shared network is applied to each descriptor, element-wise summations are utilized to generate modality-specific channel-attention maps $\mathcal{M}^{c}_{RGB}$ and $\mathcal{M}^{c}_{HHA}$. The channel-refined procedure in CMRM can be formulated as follows:

$$\begin{aligned} \mathcal{F}^{c}_{avg} &= [AvgPool(\mathcal{F}_{RGB}); AvgPool(\mathcal{F}_{HHA})] \\ &= [\mathcal{F}^{c}_{RGB\_avg}; \mathcal{F}^{c}_{HHA\_avg}], \\ \mathcal{F}^{c}_{max} &= [MaxPool(\mathcal{F}_{RGB}); MaxPool(\mathcal{F}_{HHA})] \\ &= [\mathcal{F}^{c}_{RGB\_max}; \mathcal{F}^{c}_{HHA\_max}], \end{aligned} \tag{8}$$

$$\begin{aligned} \mathcal{M}^{c}_{RGB}(\mathcal{F}_{RGB}, \mathcal{F}_{HHA}) &= \sigma(MLP_{RGB}(\mathcal{F}^{c}_{avg}) + MLP_{RGB}(\mathcal{F}^{c}_{max})) \\ \mathcal{M}^{c}_{HHA}(\mathcal{F}_{RGB}, \mathcal{F}_{HHA}) &= \sigma(MLP_{HHA}(\mathcal{F}^{c}_{avg}) + MLP_{HHA}(\mathcal{F}^{c}_{max})) \end{aligned} \tag{9}$$

where σ denotes the sigmoid function and $[;]$ denotes the concatenation. It is worth mentioning that, different from the CBAM, both MLP have weights of $\boldsymbol{W_0} \in \mathbb{R}^{C/r\times 2C}$ and $\boldsymbol{W_1} \in \mathbb{R}^{C\times 2C/r}$; r is the reduction ratio. Furthermore, $\mathcal{F}^{c}_{RGB\_avg} \in \mathbb{R}^{C\times 1\times 1}$ and $\mathcal{F}^{c}_{RGB\_max} \in \mathbb{R}^{C\times 1\times 1}$ represent the RGB descriptors, $\mathcal{F}^{c}_{HHA\_avg} \in \mathbb{R}^{C\times 1\times 1}$ and $\mathcal{F}^{c}_{HHA\_max} \in \mathbb{R}^{C\times 1\times 1}$ represent the HHA descriptors, and $\mathcal{F}^{c}_{avg} \in \mathbb{R}^{2C\times 1\times 1}$ and $\mathcal{F}^{c}_{max} \in \mathbb{R}^{2C\times 1\times 1}$ represent the cross-modality descriptors.

Following the optimization of the features by the channel-attention module, the pair of feature maps—$\mathcal{F}^{cr}_{RGB}$ and $\mathcal{F}^{cr}_{HHA}$—are fed into the spatial-attention module. Similar to the implementation of channel attention, the spatial-attention module initially aggregates cross-modality features via average-pooling and max-pooling along the channel axis, and infers four 2D maps, among which $\mathcal{F}^{s}_{RGB\_avg}$ and $\mathcal{F}^{s}_{RGB\_max}$ are generated from refined RGB

feature maps, while $\mathcal{F}^s_{HHA\_avg}$ and $\mathcal{F}^s_{HHA\_max}$ are generated from refined HHA feature maps. Then, the concatenation is applied to combine all maps. The cross-modality map is then convolved by two independent standard convolution layers with kernel size of 7×7 to generate the modality-specific spatial-attention map $\mathcal{M}^s_{RGB}$ and $\mathcal{M}^s_{HHA}$. The spatial-refined procedure in the CMRG can be formulated as follows:

$$\begin{aligned} \mathcal{F}^s &= [AvgPool(\mathcal{F}_{RGB}); MaxPool(\mathcal{F}_{RGB}); AvgPool(\mathcal{F}_{HHA}); MaxPool(\mathcal{F}_{HHA})] \\ &= [\mathcal{F}^s_{RGB\_avg}; \mathcal{F}^s_{RGB\_max}; \mathcal{F}^s_{HHA\_avg}; \mathcal{F}^s_{HHA\_max}], \end{aligned} \tag{10}$$

$$\begin{aligned} \mathcal{M}^s_{RGB}(\mathcal{F}^{cr}_{RGB}, \mathcal{F}^{cr}_{HHA}) &= \sigma(f^{7\times7}_{RGB}(\mathcal{F}^s)) \\ \mathcal{M}^c_{HHA}(\mathcal{F}^{cr}_{RGB}, \mathcal{F}^{cr}_{HHA}) &= \sigma(f^{7\times7}_{HHA}(\mathcal{F}^s)), \end{aligned} \tag{11}$$

where σ denotes the sigmoid function, $f^{7\times7}$ represents the 7×7 convolutional layer, and $[;]$ refers to the concatenation. $\mathcal{F}^s_{RGB\_avg} \in \mathbb{R}^{1\times H\times W}$ and $\mathcal{F}^s_{RGB\_max} \in \mathbb{R}^{1\times H\times W}$ represent the RGB maps, and $\mathcal{F}^s_{HHA\_avg} \in \mathbb{R}^{1\times H\times W}$ and $\mathcal{F}^s_{HHA\_max} \in \mathbb{R}^{1\times H\times W}$ represent the HHA maps.

To refine the features by taking advantages of different physical significance, the CMRG utilizes combined information from RGB and HHA to generate attention maps instead of relying solely on a single modality. Due to the limited sensing capabilities of the depth camera, this strategy improves the sturdiness and effectiveness of the backbone, especially in the HHA branch. To be specific, for the purpose of generating fine attention maps, the descriptors (maps) are derived using both average-pooling and max-pooling, with average-pooling descriptors (maps) representing the global information and max-pooling descriptors (maps) representing prominent information, thereby improving the availability of the attention maps. In addition, the CMRG employs channel- and spatial-attention mechanisms to improve the representation and aggregation of cross-modality information. In this structure, the channel-attention module is primarily responsible for capturing 'what' is important, whereas the spatial-attention module is primarily responsible for determining 'where' should be prioritized.

3.2.3. Bi-Directional Multi-Step Propagation

It should be noticed that the CMRG can properly fuse the features from both branches. Moreover, Bi-directional Multi-step Propagation (BMP) is employed to reduce model complexity and improve propagation efficiency.

BMP propagates the refined results to the next layer in the encoder for more accurate and efficient encoding of the RGB and HHA features by minimizing the fusion result to half its original values rather than adding elements directly. The procedure of the BMP can be formulated as follows:

$$\begin{aligned} RGB_{out} &= (REF + RGB_{in})/2 \\ HHA_{out} &= (REF + HHA_{in})/2 \end{aligned} \tag{12}$$

where REF denotes the output of the CMRG.

3.3. Pyramid Supervision

The pyramid-supervision training strategy mitigates the gradient vanishing issue by incorporating supervised learning over multiple levels; furthermore, it utilizes the features in different scales to improve the final semantic performance.

As illustrated in Figure 3, four intermediate side outputs are implemented, which are derived from the features of the four up-layers for pyramid supervision in addition to the final output of the decoder. Each output is generated after the corresponding feature

maps are convolved by a 1×1 convolution layer. Unlike the final output, which has the original spatial resolution, the four side outputs have a different spatial resolution, with $1/2$, $1/4$, $1/8$, and $1/16$ the height and width of the final output. The loss function of pyramid supervision is formulated as follows:

$$Loss(O_1, ...O_5) = \sum_{i=0}^{n} Loss(O_i) \tag{13}$$

where

$$Loss(O_n) = \frac{1}{N} \sum_{i} -log(\frac{exp(s_i[g_i])}{\sum_k exp(s_i[k])}) \tag{14}$$

where $Loss(O_n)$ is the loss function for the final output or the side outputs. $g_i \in \mathbb{R}$ denotes the class index on the pixel i of the groundtruth semantic map. $s_i \in \mathbb{R}^{N_c}$ denotes the vector on the pixel i of the output score map and N_c denotes the class number of the dataset.

4. Experiments

In order to verify the effectiveness of our proposed method, we conduct evaluation experiments on RGB-D public datasets NYUDv2 [20] and SUN RGB-D [21]. To evaluate the results, we compare the semantic-segmentation performance of various methods on three metrics: pixel accuracy, mean pixel accuracy, and mean intersection over union [4]. In addition, ablation experiments are performed on NYUDv2 with ResNet-50 [44] as the backbone, and certain analysis and discussion are provided.

4.1. Datasets

In this section, we introduce the public datasets that are utilized in our experiment. The two public datasets are:

NYUDv2 [20] The NYUDv2 dataset consists of 1449 indoor RGB-D images with dense pixel-wise annotation. According to the official instructions [20], we split them into 795 training images and 654 testing images. A 40-category setting is adopted as in [46].

SUN RGB-D [21] The SUN RGB-D dataset contains 10,335 indoor RGB-D images with 37 categories, which include images from NYUDv2 [20], Berkeley B3DO [47], SUN3D [48], and newly captured RGB-D images. We divide the dataset into 5285 training images and 5050 testing images according to the official setting.

4.2. Implementation Details

We implement our experiments with Python 3.8 in the Ubuntu 18 operating system with Pytorch [49] framework. All models are trained on one Nvidia RTX A5000 graphics card with batch size of 7. Additionally, we use ResNet-50 pre-trained on ImageNet [50] as the backbone of the two branches in the encoder. We adopt the SGD optimizer with momentum 0.9 and weight decay 0.0005. The initial learning rate is 0.001 and it decays by a factor of 0.8 every 100 epochs. We employ the warm-up strategy in the first 15 epochs. The network is trained for 800 epochs with the NYUDv2 and SUN RGB-D datasets. We set the reduction ratio to 8 in all attention modules.

As for the data augmentation strategy, we utilize random rotation (with 90 or 180 degree clockwise rotation), random flipping (left-right or top-bottom flip), random scaling (with [1.0, 1.6] scale), random color jittering (brightness, saturation, contrast change), and random cropping (to 480×640 pixels).

4.3. Experimental Results and Comparisons

We compare our CMANet with existing semantic-segmentation methods on the NYUDv2 and SUN RGB-D datasets. The results of the three aforementioned metrics on the two datasets are displayed in Tables 2 and 3, whereas the details of class IoU on NYUDv2 are displayed in Table 4.

As shown Table 2, our CMANet outperforms most of the state-of-the-art methods in semantic segmentation on the NYUDv2 dataset. On the three evaluation metrics, CMANet achieves 74.2% pixel accuracy, 60.6% mean accuracy, and 47.6% mean IoU. Additionally, on the most important metric—mean IoU—CMANet achieves a 2.7% improvement compared to RefineNet-101 [45], 1.7% compared to LSD-GF [51], and 0.1% compared to RAFNet [43]. It is noticed that we only utilize ResNet-50 as our backbone, suggesting that CMANet is capable of better performance with a more powerful backbone. However, the performance of CMANet is lower than that of RDFNet on mIoU by 0.1%. Despite its slight deficiency in segmentation performance, CMANet displays an improvement in memory and computing complexity according to the model efficiency analysis. Furthermore, additional experiments are conducted to compare the utilization of HHA images and depth images. The results display that the application of HHA images can slightly improve the performance of CMANet, with a 0.3% increase on mIoU.

Table 2. Comparison of CMANet with state-of-the-arts methods on the NYUDv2 dataset in 40 classes. The results are reported in terms of percentage (%) of pixel accuracy, mean accuracy, and mean IoU. The results of other methods originate from the corresponding citation.

Method	Data	Pixel Acc. (%)	Mean Acc. (%)	Mean IoU. (%)
Gupta et al. [46]	RGB-D	60.3	35.1	28.6
Deng et al. [41]	RGB-D	63.8	-	31.5
FCN-16s [4]	RGB-HHA	65.4	46.1	34.0
Wang et al. [14]	RGB-D	-	47.3	-
Context [42]	RGB	70.0	53.6	40.6
STD2P [52]	RGB-D	70.1	53.8	40.1
3DGNN [53]	RGB-HHA	-	55.7	43.1
Depth-aware [54]	RGB-HHA	-	56.3	43.9
LSD-GF [51]	RGB-HHA	71.9	60.7	45.9
RefineNet-101 [45]	RGB	72.8	57.8	44.9
RDFNet-50 [13]	RGB-HHA	74.8	60.4	47.7
RAFNet-50 [43]	RGB-D	73.8	60.3	47.5
CMANet-50-depth (ours)	RGB-D	73.9	59.8	47.3
CMANet-50 (ours)	RGB-HHA	74.2	60.2	47.6

As shown in Table 4, we also compare the category-wise results on class IoU. CMANet performs better than RefineNet-101 and LSD-GF over 28 and 22 classes (40 classes in total), respectively. These results demonstrate the robustness and effectiveness of CMANet in indoor-scene semantic segmentation.

Table 3. Comparison of CMANet with state-of-the-art methods on the SUN RGB-D dataset in 37 classes. The results are reported in terms of percentage (%) of pixel accuracy, mean accuracy, and mean IoU.

Method	Data	Pixel Acc. (%)	Mean Acc. (%)	Mean IoU. (%)
SegNet [5]	RGB	71.2	45.9	30.4
FuseNet [19]	RGB-D	76.3	48.3	37.3
Depth-aware [54]	RGB-HHA	-	53.5	42.0
Context [42]	RGB	78.4	53.4	42.3
3DGNN [53]	RGB-D	-	57.0	45.9
RefineNet-101 [45]	RGB	80.4	57.8	45.7
RedNet-34 [18]	RGB-D	80.8	58.3	46.8
CMANet-50 (ours)	RGB-HHA	81.1	59.3	47.2

Due to the limited data scale of NYUDv2 dataset, we also compare the semantic-segmentation performance of CMANet on the large-scale SUN RGB-D dataset with other existing methods, following the same training and testing strategy as on the NYUDv2 dataset. The comparison results are displayed in Table 3, where CMANet achieves the best perfor-

mance among all the methods on all three evaluation metrics. The semantic-segmentation performance of CMANet on the SUN RGB-D dataset further verifies its validity.

Table 4. Comparison of CMANet with state-of-the-art methods on the NYUDv2 dataset in 40 classes. The results are reported in terms of percentage (%) of IoU.

Method	**Wall**	**Floor**	**Cabinet**	**Bed**	**Chair**	**Sofa**	**Table**	**Door**	**Window**	**Bookshelf**
Gupta et al. [46]	68.0	81.3	44.9	65.0	47.9	29.9	20.3	32.6	9.0	18.1
Deng et al. [41]	65.6	79.2	51.9	66.7	41.0	55.7	36.5	20.3	33.2	32.6
FCN-16s [4]	69.9	79.4	50.3	66.0	47.5	53.2	32.8	22.1	39.0	36.1
STD2P [52]	72.7	85.7	55.4	73.6	58.5	60.1	42.7	30.2	42.1	41.9
LSD-GF [51]	78.5	87.1	56.6	70.1	65.2	63.9	46.9	35.9	47.1	48.9
RefineNet-101 [45]	77.5	82.9	58.7	65.7	59.1	57.8	40.1	36.7	45.8	42.8
CMANet-50 (ours)	77.7	86.2	59.6	72.5	60.3	61.1	43.3	35.5	43.8	38.6
Method	**Picture**	**Counter**	**Blind**	**Desk**	**Shelf**	**Curtain**	**Dresser**	**Pillow**	**Mirror**	**Mat**
Gupta et al. [46]	40.3	51.3	42.0	11.3	3.5	29.1	34.8	34.4	16.4	28.0
Deng et al. [41]	44.6	53.6	49.1	10.8	9.1	47.6	27.6	42.5	30.2	32.7
FCN-16s [4]	50.5	54.2	45.8	11.9	8.6	32.5	31.0	37.5	22.4	13.6
STD2P [52]	52.9	59.7	46.7	13.5	9.4	40.7	44.1	42.0	34.5	35.6
LSD-GF [51]	54.3	66.3	51.7	20.6	13.7	49.8	43.2	50.4	48.5	32.2
RefineNet-101 [45]	60.1	56.8	61.4	22.6	12.3	53.5	38.3	39.6	38.7	29.7
CMANet-50 (ours)	60.9	62.5	56.1	21.7	10.0	56.1	50.1	46.4	45.8	37.2
Method	**Cloths**	**Ceiling**	**Books**	**Refridg**	**TV**	**Paper**	**Towel**	**Shower**	**Box**	**Board**
Gupta et al. [46]	4.7	60.5	6.4	14.5	31.0	14.3	16.3	4.2	2.1	14.2
Deng et al. [41]	12.6	56.7	8.9	21.6	19.2	28.0	28.6	22.9	1.6	1.0
FCN-16s [4]	18.3	59.1	27.3	27.0	41.9	15.9	26.1	14.1	6.5	12.9
STD2P [52]	22.2	55.9	29.8	41.7	52.5	21.1	34.4	15.5	7.8	29.2
LSD-GF [51]	24.7	62.0	34.2	45.3	53.4	27.7	42.6	23.9	11.2	58.8
RefineNet-101 [45]	24.4	66.0	33.0	52.4	52.6	31.3	36.8	23.6	11.1	63.7
CMANet-50 (ours)	21.1	75.3	33.1	55.1	63.3	30.1	40.1	32.1	14.3	62.5
Method	**Person**	**Stand**	**Toilet**	**Sink**	**Lamp**	**Bathtub**	**Bag**	**Othstr**	**Othfurn**	**Otherprop**
Gupta et al. [46]	0.2	27.2	55.1	37.5	34.8	38.2	0.2	7.1	6.1	23.1
Deng et al. [41]	9.6	30.6	48.4	41.8	28.1	27.6	0.0	9.8	7.6	24.5
FCN-16s [4]	57.6	30.1	61.3	44.8	32.1	39.2	4.8	15.2	7.7	30.0
STD2P [52]	60.7	42.2	62.7	47.4	38.6	28.5	7.3	18.8	15.1	31.4
LSD-GF [51]	53.2	54.1	80.4	59.2	45.5	52.6	15.9	12.7	16.4	29.3
RefineNet-101 [45]	78.6	38.6	68.4	53.2	45.9	32.9	14.6	32.9	18.7	36.4
CMANet-50 (ours)	77.3	40.8	70.9	58.9	47.9	57.3	13.6	31.2	19.1	38.1

4.4. Ablation Study

In order to investigate the functionality of the proposed network and its processing modules, extensive ablation experiments are performed on the NYUDv2 dataset. Each experiment is conducted with the same hyper-parameter settings during training and testing periods.

The ablation study w.r.t. CMRGs is performed to verify the functionality of CMRGs in different encoding stages. As displayed in Table 5, among the first three defective models, each of the first four defective models removes certain CMRGs, but the fifth contains them all, i.e., the original CMANet. It is interesting to find that the second defective models (G_3, G_4 and G_5 are removed) outperform the first one (G_1, G_2 are removed). Furthermore, with the gradual stacking of the CMRG from lower to higher stages, the performances of the defective models become better, with the best performance using the CMRG in all stages. From these facts, it is clear that the cross-modality fusion, i.e., the utilization of CMRGs, plays a vital role in performance improvement, especially in the earlier stages. This can be recognized by the fact that the low-level features are rough and compatible in a CNN-based model. The original CMANet performs the best, proving that the multi-stage cross-modality fusion is effective and the CMRGs are mutually reinforcing.

Table 5. Ablation study of CMRGs on NYUDv2 dataset in 40 classes. The results are reported in terms of percentage (%) of mean IoU. *G* denotes the CMRG in different encoding stage. The best performance is marked in bold.

Backbone	G_1	G_2	G_3	G_4	G_5	Mean IoU. (%)
ResNet-50			✓	✓	✓	46.5
ResNet-50	✓	✓				46.7
ResNet-50	✓	✓	✓			47.2
ResNet-50	✓	✓	✓	✓		46.8
ResNet-50	✓	✓	✓	✓	✓	**47.6**

Additionally, we also conduct an ablation study on CMRGs, skip connections, and pyramid supervision to evaluate the effect of these strategies; the results are displayed in Table 6. The first defective model is the baseline without any strategy; the second, third, and fourth defective models remove one corresponding strategy at a time; the fifth is the original model. We compare them on mean accuracy and mean IoU, showing that the order of influence from most significant to most minor is CMRGs, skip connections, pyramid supervision. According to the results, CMRGs can effectively improve the performance of semantic segmentation, while skip connections and pyramid supervision provide slight improvements.

Table 6. Ablation study of CMRGs, skip connections, and pyramid supervision on NYUDv2 dataset in 40 classes. The results are reported in terms of percentage (%) of mean IoU. The best performance is marked in bold.

Backbone	Baseline	CMRG	Skip	Pyramid	Mean IoU. (%)
ResNet-50	✓				44.2
ResNet-50	✓	✓		✓	46.8
ResNet-50	✓		✓	✓	45.9
ResNet-50	✓	✓	✓		47.0
ResNet-50	✓	✓	✓	✓	**47.6**

4.5. Model Efficiency Analysis

Complexity in terms of time and space is an important metric to evaluate the efficiency of the model. Accordingly, we compare our CMANet with [13,53] to verify the model's efficiency. According to the results displayed in Table 7, our method achieves 23.8% parameters reduction and 19.4% FLOPs reduction compared to RDFNet [13] while maintaining almost the same performance. Meanwhile, CMANet outperforms 3DGNN [53] with a 20.5% inference time reduction and a 4.5% mean IoU improvement. However, CMANet has a larger number of parameters than 3DGNN and a lower mean IoU than RDFNet-50. Consequently, it can be seen that CMANet achieves a balance between model complexity and accuracy, which will be further improved in our future research.

Table 7. Comparison of model complexity. The results are reported in terms of the number of parameters (million), the computing complexity of FLOPs (gigabyte), inference time (millisecond), and mean IoU (%). The inference time and FLOPs are evaluated on an Nvidia 1660Ti GPU with the RGB input of $3 \times 480 \times 640$ and the HHA input of $3 \times 480 \times 640$.

Method	Backbone	Params. (M)	FLOPs. (G)	Inference Time. (ms)	Mean IoU. (%)
3DGNN [53]	ResNet-101	47.3	-	492.5	43.1
RedNet-101 [18]	ResNet-101	121.2	-	268.5	-
RefineNet-101 [45]	ResNet-101	126.0	-	248.4	44.9
RDFNet-50 [13]	ResNet-50	153.3	168.9	368.2	47.7
CMANet-50 (ours)	ResNet-50	117.8	137.2	391.5	47.6

4.6. Visualization

For the purpose of evaluating the performance of semantic segmentation the analysis is supposed to be not only quantitative but also qualitative; we should pay more attention to the interpretability of the results. Therefore, we conduct visualizations of the CMANet results.

Semantic Segmentation Qualitative Visual Results In Figure 8, we visualize some typical examples of semantic segmentation with our baseline, the defective models, and our proposed method CMANet. In Figure 8a, the bedroom has few objects on the bed, whereas Figure 8b has disorganized items on the bed. Figure 8c represents the hallway scene with a complex structure. Figure 8d has an obvious lighting imbalance in the bedroom. In Figure 8e, the table is cluttered with small and numerous objects. Figure 8f presents a scene in which there are not only strong lighting conditions, but also many overlapped and similar-texture objects. Compared to other methods, CMANet promotes semantic-segmentation results from the perspective of details and the misclassification phenomenon.

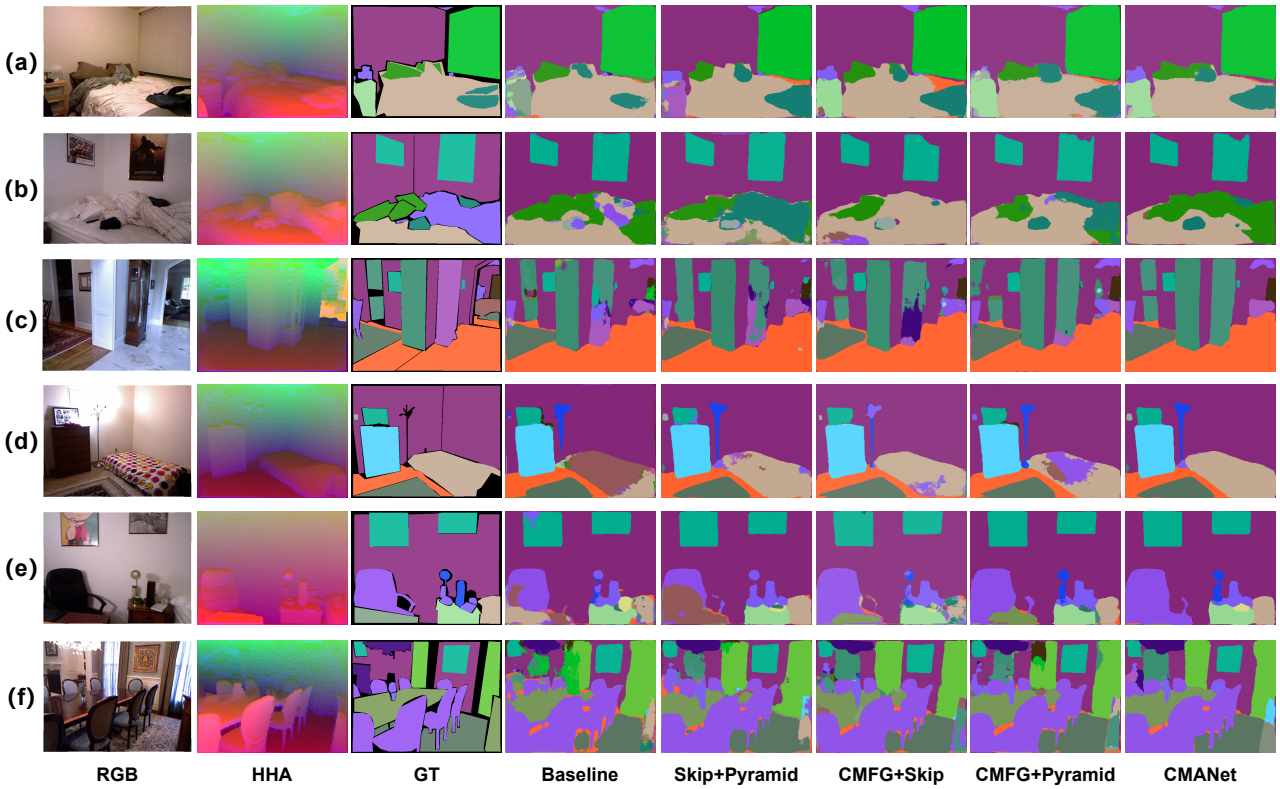

Figure 8. Semantic segmentation qualitative visual results on NYUDv2 dataset. The (**a**) represents the bedroom has few objects on the bed; (**b**) has disorganized items on the bed; (**c**) represents the hallway scene with a complex structure; (**d**) has an obvious lighting imbalance in the bedroom; (**e**) represents the table which is cluttered with small and numerous objects; (**f**) presents a scene in which there are not only strong lighting conditions, but also many overlapped and similar-texture objects.

Pyramid Supervision Visualization We conduct visualizations of the pyramid supervision by generating semantic maps from the final output and four side outputs; meanwhile, the performance of some randomly sampled examples is illustrated in Figure 9. The fourth to eighth columns denote the final output, and the four side outputs have spatial resolutions of 1/2, 1/4, 1/8, and 1/16, respectively, the height and width of the final output. As the semantic-segmentation of the outputs progresses from right to left, the refinement of pixel classification gradually improves. However, the results demonstrate that the output with a lower spatial resolution has a more remarkable performance on large-object (wall, TV, etc.)

segmentation, as well as edge extraction, owing to its larger receptive field. In this way, the supervision in low-spatial-resolution outputs assists the higher ones via the recognition of boundaries and large objects; meanwhile, the supervision in high-spatial-resolution outputs can refine the semantic information. As a result, the pyramid supervision enables the enhancement of the final result via multi-scale semantic analysis.

Channel Attention Visualization To verify the effectiveness of the channel-attention mechanism, which can enhance the advantages and filter the drawback, we conduct visualizations of the RGB and HHA channel-refined features. In Figure 10, we randomly select two high-weighting feature maps from the channel-refined features. As shown in Figure 10a, the RGB feature maps focus on the significant texture regions (e.g., the wall hangings, the curtains), while the HHA feature maps are mainly concerned with the significant structure regions (e.g., the office chair, the bookshelf), as illustrated in Figure 10b. The results demonstrate that the CMRG enhances feature extraction by paying attention to essential features.

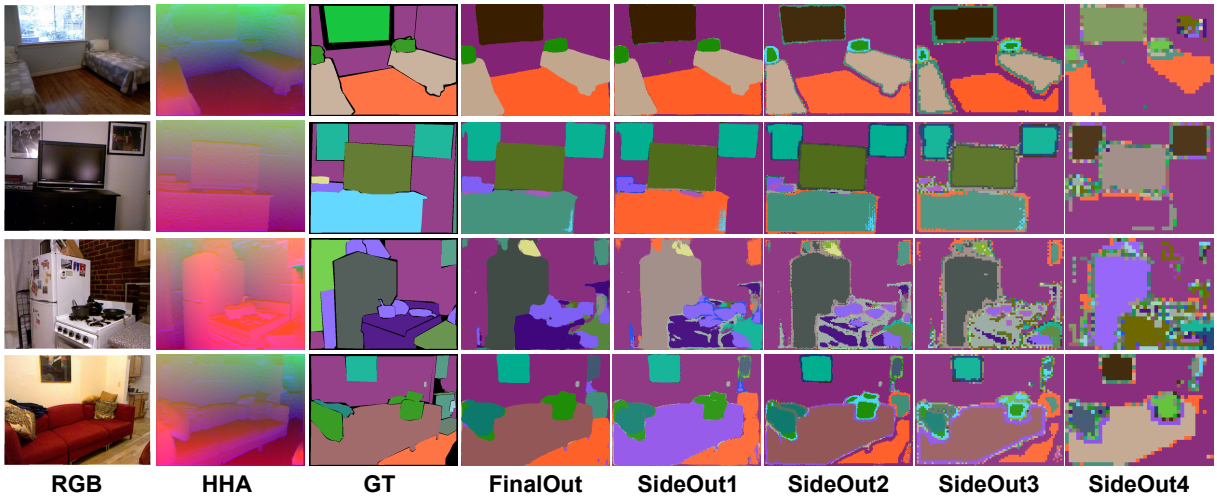

Figure 9. Visualization of pyramid supervision.

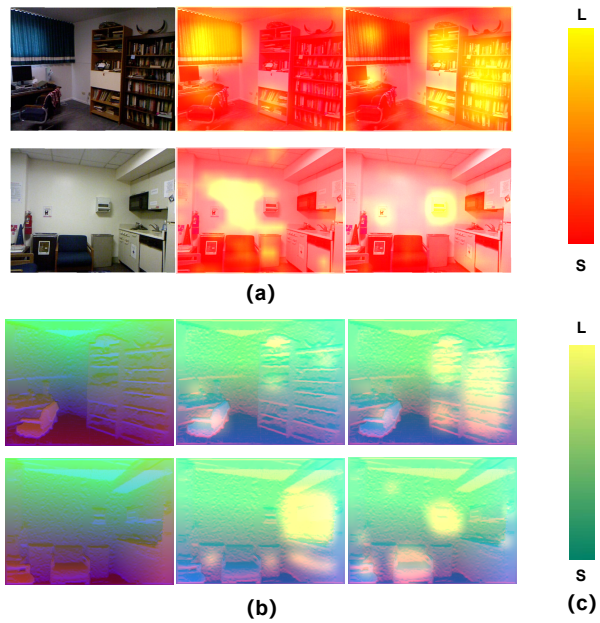

Figure 10. Visualization of channel attention. (**a**) The visual results on RGB channel-refined features; (**b**) the visual results on HHA channel-refined features; (**c**) the color bar.

Cross-Modality Refine Gate Visualization In order to understand refinement by the CMRG, we visualize the output of the CMRG-refined features. As illustrated in Figure 11, we conduct visualizations on some typical feature maps on two random sampled examples. In the first row of Figure 11, the CMRG refinement enhances the regions that share the same labels, such as the sofa, the floor, or the wall, while in the second row, the refinement emphasizes the table and objects on the table. As a result, the CMRG is capable of effectively building connections among the regions with similar characteristics, even if they are geographically dispersed.

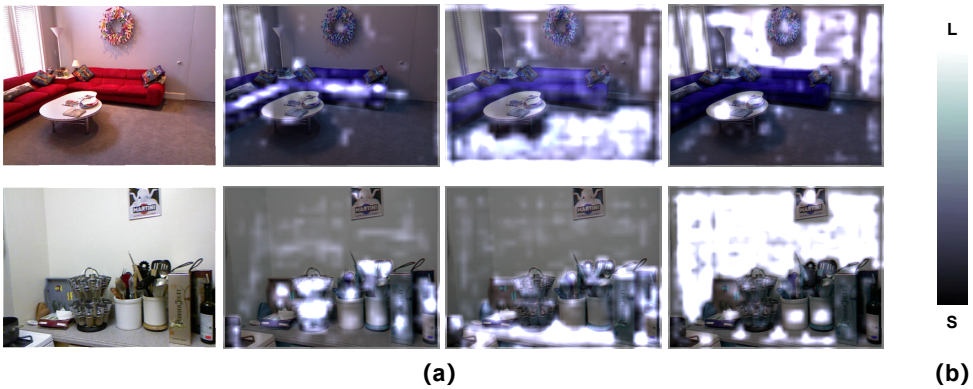

Figure 11. Visualization of the cross-modality refine gate. (**a**) Visual results of the output of the CMRG; (**b**) the color bar.

5. Conclusions

In this paper, we proposed a novel CMANet method for indoor-scene semantic segmentation, which utilizes HHA and RGB images to enhance the robustness of segmentation in indoor scenes. According to our experiments, CMANet not only facilitates the learning process via the enhancement of the representation, robustness, and discrimination of the feature embedding, but also takes the advantage of cross-modality features. CMANet employs the encoder–decoder architecture. The encoder has a two-parallel-branch backbone that can extract and aggregate the specific features from RGB and HHA; meanwhile, the decoder generates multi-scale semantic maps that can improve the final segmentation results. Specifically, we designed the CMRG, which is the most crucial component in CMANet. The CMRG employs a sequence of cross-modality channel- and spatial-attention modules. The channel-attention module is responsible for capturing 'what' is important, whereas the spatial-attention module is responsible for determining 'where' should be prioritized. The CMRG filters the noisy information and integrates the features from different modalities (RGB and HHA). The CMRG can effectively enhance representations from both modalities by selecting key features and establishing connections between relevant regions. Additionally, we employ a bi-directional multi-step propagation strategy to provide assistance in propagating. The results of the ablation study and visualization demonstrate the significance of each proposed component. The experiments on the NYUDv2 and SUN RGB-D datasets verify the robustness and effectiveness of CMANet, and the results illustrate that the network outperforms the existing indoor-scene semantic-segmentation methods and achieves a new state-of-art performance. In our future research, we will focus more on increasing the efficiency of our network by reducing time and space complexity. Moreover, we will consider applying semi-supervised or weakly-supervised learning strategies for indoor-scene semantic segmentation due to the limited dataset scale and inaccurate data labeling.

Author Contributions: Methodology, Z.K. and L.Z.; data processing and experimental results analysis, L.Z.; funding acquisition, Z.K.; supervision and suggestions, Z.K., M.Z., Z.W. and X.Y.; writing—review and editing, Z.K., L.Z. and X.Y.; visualization, L.Z.; investigation, Z.C. and C.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (Grant No. 41872207), and the Zhejiang Provincial Natural Science Foundation (Grant No. LY20F030018).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

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

References

1. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
2. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Diego, CA, USA, 20–25 June 2005; Voluem 1, pp. 886–893.
3. Ren, X.; Bo, L.; Fox, D. Rgb-(d) scene labeling: Features and algorithms. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 2759–2766.
4. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
5. Badrinarayanan, V.; Kendall, A.; Cipolla, R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [CrossRef] [PubMed]
6. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 834–848. [CrossRef] [PubMed]
7. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention; Springer: Munich, Germany, 5–9 October 2015; pp. 234–241.
8. Ge, P.; Chen, Y.; Wang, G.; Weng, G. An active contour model driven by adaptive local pre-fitting energy function based on Jeffreys divergence for image segmentation. *Expert Syst. Appl.* **2022**, *210*, 118493. [CrossRef]
9. Ge, P.; Chen, Y.; Wang, G.; Weng, G. A hybrid active contour model based on pre-fitting energy and adaptive functions for fast image segmentation. *Pattern Recognit. Lett.* **2022**, *158*, 71–79. [CrossRef]
10. Wang, Z.; Li, T.; Pan, L.; Kang, Z. Scene semantic segmentation from indoor RGB-D images using encode-decoder fully convolutional networks. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2017**, *XLII-2/W7*, 397–404. [CrossRef]
11. Gupta, S.; Girshick, R.; Arbeláez, P.; Malik, J. Learning rich features from RGB-D images for object detection and segmentation. In Proceedings of the European Conference on Computer Vision; Springer: Zurich, Switzerland, 6–12 September 2014; pp. 345–360.
12. Chen, X.; Lin, K.Y.; Wang, J.; Wu, W.; Qian, C.; Li, H.; Zeng, G. Bi-directional cross-modality feature propagation with separation-and-aggregation gate for RGB-D semantic segmentation. In Proceedings of the European Conference on Computer Vision; Springer: Glasgow, UK, 23–28 August 2020; pp. 561–577.
13. Park, S.J.; Hong, K.S.; Lee, S. Rdfnet: Rgb-d multi-level residual feature fusion for indoor semantic segmentation. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 4980–4989.
14. Wang, J.; Wang, Z.; Tao, D.; See, S.; Wang, G. Learning common and specific features for RGB-D semantic segmentation with deconvolutional networks. In Proceedings of the European Conference on Computer Vision; Springer: Amsterdam, The Netherlands, 11–14 October 2016; pp. 664–679.
15. Zhou, H.; Qi, L.; Huang, H.; Yang, X.; Wan, Z.; Wen, X. CANet: Co-attention network for RGB-D semantic segmentation. *Pattern Recognit.* **2022**, *124*, 108468. [CrossRef]
16. Farabet, C.; Couprie, C.; Najman, L.; LeCun, Y. Learning hierarchical features for scene labeling. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *35*, 1915–1929. [CrossRef] [PubMed]
17. Couprie, C.; Farabet, C.; Najman, L.; LeCun, Y. Indoor semantic segmentation using depth information. *arXiv* **2013**, arXiv:1301.3572.
18. Jiang, J.; Zheng, L.; Luo, F.; Zhang, Z. Rednet: Residual encoder-decoder network for indoor rgb-d semantic segmentation. *arXiv* **2018**, arXiv:1806.01054.
19. Hazirbas, C.; Ma, L.; Domokos, C.; Cremers, D. Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture. In Proceedings of the Asian Conference on Computer Vision; Springer: Taipei, Taiwan, 20–24 November 2016; pp. 213–228.
20. Silberman, N.; Hoiem, D.; Kohli, P.; Fergus, R. Indoor segmentation and support inference from rgbd images. In Proceedings of the European Conference on Computer Vision; Springer: Florence, Italy, 7–13 October 2012; pp. 746–760.
21. Song, S.; Lichtenberg, S.P.; Xiao, J. Sun rgb-d: A rgb-d scene understanding benchmark suite. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 567–576.
22. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the 30th Annual Conference on Neural Information Processing System, Long Beach, CA, USA, 4–9 December 2017.

23. Lin, Z.; Feng, M.; Santos, C.N.d.; Yu, M.; Xiang, B.; Zhou, B.; Bengio, Y. A structured self-attentive sentence embedding. *arXiv* **2017**, arXiv:1703.03130.
24. Pavlopoulos, J.; Malakasiotis, P.; Androutsopoulos, I. Deeper attention to abusive user content moderation. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, 9–11 September 2017; pp. 1125–1135.
25. Zhang, Q.; Shi, Y.; Zhang, X. Attention and boundary guided salient object detection. *Pattern Recognit.* **2020**, *107*, 107484. [CrossRef]
26. Wang, D.; Xiang, S.; Zhou, Y.; Mu, J.; Zhou, H.; Irampaye, R. Multiple-Attention Mechanism Network for Semantic Segmentation. *Sensors* **2022**, *22*, 4477. [CrossRef] [PubMed]
27. Chen, L.; Zhang, H.; Xiao, J.; Nie, L.; Shao, J.; Liu, W.; Chua, T.S. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 5659–5667.
28. Zhang, H.; Dana, K.; Shi, J.; Zhang, Z.; Wang, X.; Tyagi, A.; Agrawal, A. Context encoding for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7151–7160.
29. Lee, H.; Kim, H.E.; Nam, H. Srm: A style-based recalibration module for convolutional neural networks. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 1854–1862.
30. Yang, Z.; Zhu, L.; Wu, Y.; Yang, Y. Gated channel transformation for visual recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 11794–11803.
31. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
32. Gao, Z.; Xie, J.; Wang, Q.; Li, P. Global second-order pooling convolutional networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3024–3033.
33. Li, X.; Wang, W.; Hu, X.; Yang, J. Selective kernel networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 510–519.
34. Mnih, V.; Heess, N.; Graves, A.; Kavukcuoglu, K. Recurrent models of visual attention. In Proceedings of the 27th Annual Conference on Neural Information Processing System, Montreal, QC, Canada, 8–13 December 2014.
35. Ba, J.; Mnih, V.; Kavukcuoglu, K. Multiple object recognition with visual attention. *arXiv* **2014**, arXiv:1412.7755.
36. Oktay, O.; Schlemper, J.; Folgoc, L.L.; Lee, M.; Heinrich, M.; Misawa, K.; Mori, K.; McDonagh, S.; Hammerla, N.Y.; Kainz, B.; et al. Attention u-net: Learning where to look for the pancreas. *arXiv* **2018**, arXiv:1804.03999.
37. Zhao, H.; Zhang, Y.; Liu, S.; Shi, J.; Loy, C.C.; Lin, D.; Jia, J. Psanet: Point-wise spatial attention network for scene parsing. In Proceedings of the European Conference on Computer Vision (ECCV); Springer: Munich, Germany, 8–14 September 2018; pp. 267–283.
38. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16 $\times$ 16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
39. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
40. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV); Springer: Munich, Germany, 8–14 September 2018; pp. 3–19.
41. Deng, Z.; Todorovic, S.; Jan Latecki, L. Semantic segmentation of rgbd images with mutex constraints. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1733–1741.
42. Lin, G.; Shen, C.; Van Den Hengel, A.; Reid, I. Efficient piecewise training of deep structured models for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 3194–3203.
43. Yan, X.; Hou, S.; Karim, A.; Jia, W. RAFNet: RGB-D attention feature fusion network for indoor semantic segmentation. *Displays* **2021**, *70*, 102082. [CrossRef]
44. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
45. Lin, G.; Milan, A.; Shen, C.; Reid, I. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1925–1934.
46. Gupta, S.; Arbelaez, P.; Malik, J. Perceptual organization and recognition of indoor scenes from RGB-D images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 564–571.
47. Janoch, A.; Karayev, S.; Jia, Y.; Barron, J.T.; Fritz, M.; Saenko, K.; Darrell, T. A category-level 3d object dataset: Putting the kinect to work. In *Consumer Depth Cameras for Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2013; pp. 141–165.
48. Xiao, J.; Owens, A.; Torralba, A. Sun3d: A database of big spaces reconstructed using sfm and object labels. In Proceedings of the IEEE International Conference on Computer Vision, Sydney, NSW, Australia, 1–8 December 2013; pp. 1625–1632.

49. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. Pytorch: An imperative style, high-performance deep learning library. In Proceedings of the 32th Annual Conference on Neural Information Processing System, Vancouver, BC, Canada, 8–14 December 2019.
50. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255.
51. Cheng, Y.; Cai, R.; Li, Z.; Zhao, X.; Huang, K. Locality-sensitive deconvolution networks with gated fusion for rgb-d indoor semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 3029–3037.
52. He, Y.; Chiu, W.C.; Keuper, M.; Fritz, M. Std2p: Rgbd semantic segmentation using spatio-temporal data-driven pooling. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4837–4846.
53. Qi, X.; Liao, R.; Jia, J.; Fidler, S.; Urtasun, R. 3d graph neural networks for rgbd semantic segmentation. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 5199–5208.
54. Wang, W.; Neumann, U. Depth-aware cnn for rgb-d segmentation. In Proceedings of the European Conference on Computer Vision (ECCV); Springer: Munich, Germany, 8–14 September 2018; pp. 135–150.