frontiers | Frontiers in Physics

TYPE Original Research
PUBLISHED 31 May 2024
DOI 10.3389/fphy.2024.1411559

Check for updates

OPEN ACCESS

EDITED BY
Jicheng Wang,
Jiangnan University, China

REVIEWED BY
Zhiqin Zhu,
Chongqing University of Posts and Telecommunications, China
Yu Liu,
Hefei University of Technology, China

*CORRESPONDENCE
Minghong Xie,
✉ minghongxie@163.com

RECEIVED 03 April 2024
ACCEPTED 14 May 2024
PUBLISHED 31 May 2024

CITATION
Zhang S and Xie M (2024), MIPANet: optimizing RGB-D semantic segmentation through multi-modal interaction and pooling attention.
Front. Phys. 12:1411559.
doi: 10.3389/fphy.2024.1411559

MIPANet: optimizing RGB-D semantic segmentation through multi-modal interaction and pooling attention

Shuai Zhang and Minghong Xie*

Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, China

The semantic segmentation of RGB-D images involves understanding objects appearances and spatial relationships within a scene, which necessitates careful consideration of multiple factors. In indoor scenes, the presence of diverse and disorderly objects, coupled with illumination variations and the influence of adjacent objects, can easily result in misclassifications of pixels, consequently affecting the outcome of semantic segmentation. We propose a Multi-modal Interaction and Pooling Attention Network (MIPANet) in response to these challenges. This network is designed to exploit the interactive synergy between RGB and depth modalities, aiming to enhance the utilization of complementary information and improve segmentation accuracy. Specifically, we incorporate a Multi-modal Interaction Module (MIM) into the deepest layers of the network. This module is engineered to facilitate the fusion of RGB and depth information, allowing for mutual enhancement and correction. Moreover, we introduce a Pooling Attention Module (PAM) at various stages of the encoder to enhance the features extracted by the network. The outputs of the PAMs at different stages are selectively integrated into the decoder through a refinement module to improve semantic segmentation performance. Experimental results demonstrate that MIPANet outperforms existing methods on two indoor scene datasets, NYU-Depth V2 and SUN-RGBD, by optimizing the insufficient information interaction between different modalities in RGB-D semantic segmentation. The source codes are available at https://github.com/2295104718/MIPANet.

KEYWORDS

RGB-D semantic segmentation, attention mechanism, feature fusion, multi-modal interaction, feature enhancement

1 Introduction

In recent years, Convolutional Neural Networks (CNN) have been widely used in image semantic segmentation, and more and more high-performance models have gradually replaced the traditional semantic segmentation methods. With the introduction of Fully Convolutional Neural Networks (FCN) [1, 2], which has shown great potential in semantic segmentation tasks, many researchers have proposed improved semantic segmentation models based on this way.

The advent of depth sensors and cameras [3] has expanded image research from RGB colour images to RGB-Depth (RGB-D) images, which include depth information. RGB images provide details of object appearance, such as colour and texture, while depth images

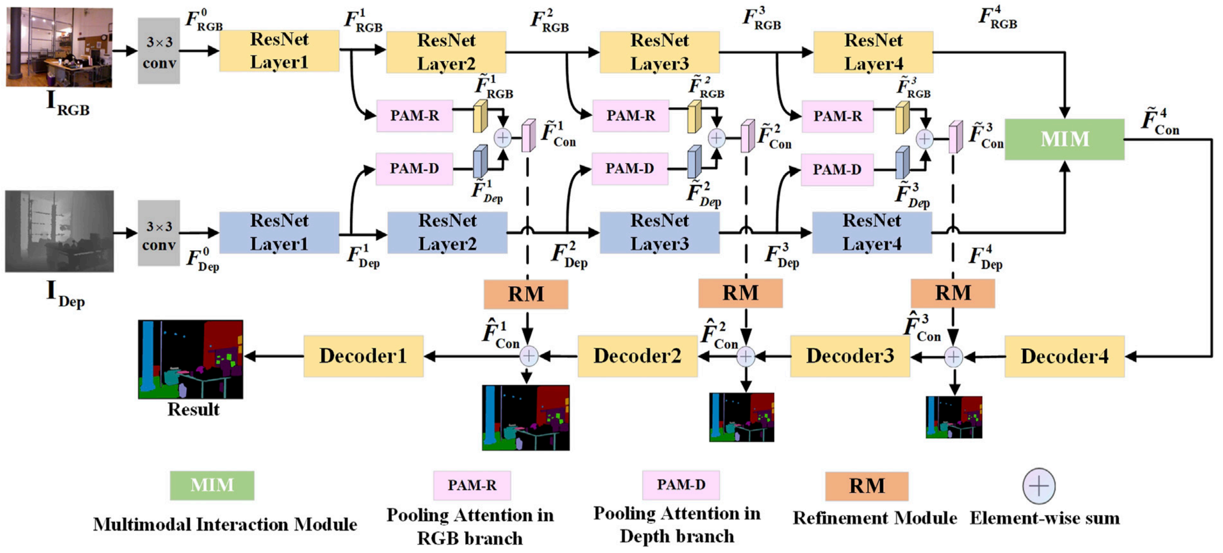


FIGURE 1
Overall architecture of the proposed network is outlined as follows: Each PAM-R or PAM-D across various levels of the network shares an identical configuration but implements distinct operations on two separate branches, yielding RGB and depth features. There are represented as $\tilde{F}^n_{RGB}$ and $\tilde{F}^n_{Dep}$. After performing an element-wise sum, we obtain $\tilde{F}^n_{Con}$, where n indicating the network level. The MIM processes RGB and depth features obtained from the ResNetLayer4 and integrates the fusion result $\tilde{F}^4_{Con}$ into the decoder.

contribute essential three-dimensional geometry information absent in RGB images, which is particularly valuable for indoor scene analysis. The fusion of these two modalities of image information would contribute to enhancing the accuracy of indoor scene semantic segmentations. [4, 5] directly concatenated RGB and depth features to create a four-channel input, resulting in improved semantic segmentation accuracy. [6] converted depth images into three channels to an HHA image which consisted of the horizontal disparity, height above ground, and angle of surface normals. Subsequently, RGB features and HHA features are fed into parallel CNNs to predict probability maps for two separate semantic segmentation. These feature maps were then fused in the final layer of the network to produce the ultimate segmentation result. Park et al. [7] and Lee et al.[8] fused the RGB features and depth features through a concatenation process. Eigen et al. [9] and Wang et al. [10] merged the RGB and depth features through directly summation. These methods fail to fully utilize the complementary information between modalities by simply summing or concatenating RGB and depth features. Shu, Li and Bai et al. [11-15] mapped text and image data to a common hash space and facilitated the interaction of information between text and images, which enhanced the performance of cross-modal retrieval. Yang et al. [16] adopted different enhancement mechanisms for RGB data and depth data, including pixel difference convolution techniques, to more effectively handle depth information. Zhao et al. [17] proposed to coordinate attention and cross-modal attention mechanisms, achieving efficient fusion of RGB and depth features and enhancing cross-modal information exchange. Yang et al. [18] developed a dual-stage refinement network (DRNet). In the initial stage, the network focuses on rough localization and feature extraction, while in the advanced stage, it concentrates on feature refinement and precise segmentation. This architecture enables more effective object boundary recovery and definition in complex scenes, thereby improving the accuracy of semantic segmentation. These methods are more effective. However, they use similar or identical operations for extracting RGB and depth features, which does not fully consider the modal differences between RGB and depth images. Moreover, they overlook the interaction between modalities, failing to maximize the complementary nature of the information from different modalities.

To solve the above problems, we propose a Multi-modal Interaction and Pooling Attention Network (MIPANet) for RGB-D semantic segmentation of indoor scenes, as illustrated in Figure 1. The proposed network adopts an encoder-decoder architecture, including two innovative modules: the Multi-modal Interaction Module (MIM) and the Pooling Attention Module (PAM). The encoder is composed of two identical CNN branches, used for extracting RGB features and depth features, respectively. In this study, RGB and depth features are incrementally extracted and fused across various network levels, utilizing spatial disparities and semantic correlations between multimodal features to optimize semantic segmentation results. In the PAM, we employ different feature enhancement strategies for RGB features and depth features. For RGB features, we use global average pooling to make the network focus on the spatial location information of RGB features. For depth features, we employ a two-step pooling operation to replace the global average pooling, aiming to guide the network during depth feature extraction to focus on the most salient parts in each channel. This allows the network to emphasize feature channels containing contours, edge information, and others, thereby enhancing feature representation. Meanwhile, it enables flexible adjustment of the output size and mitigates the impact of large outliers on the results. In the MIM, through cross-modal attention, we enable the RGB features and depth features to learn different information from each other, thereby reducing the disparity between the two modalities and enhancing their interaction. In the upsampling stage, we design a refinement module (RM) to refine the output of the PAM. This operation enriches the information of the fused features, thereby improving the accuracy of segmentation. The main contributions of this work can be summarized as follows:

(1) We propose an end-to-end multi-modal fusion network, MIPANet, incorporating multi-modal interaction and pooling attention. This method significantly enhances the feature representation of RGB and depth features, effectively focusing on regions with adjacent objects and object overlap regions in the image. Moreover, the proposed method enhances the interaction between RGB and depth features, reduces the feature disparity between modalities, enriches the fused features, and improves semantic segmentation performance.
(2) We design the MIM and PAM. Within the MIM, a cross-modal feature interaction and fusion mechanism is developed. RGB and depth features are collaboratively optimized using attention maps to extract partially detailed features. In addition, the PAM augments the extraction of RGB and depth features through distinct operations, acting as an essential supplement of information in the decoder. It facilitates feature upsampling and restoration via the RM module, ensuring a comprehensive enhancement and integration of critical details for accurate segmentation.
(3) Experimental results confirm the effectiveness of our proposed RGB-D semantic segmentation network in accurately handling indoor images in complex scenarios. The proposed model demonstrates superior semantic segmentation performance compared to other methods on the publicly available NYU-Depth V2 and SUN RGB-D datasets. The visualization results demonstrate that our method focuses more effectively on regions of the image where neighbouring objects may be similar and overlap between objects, resulting in more accurate segmentation outcomes in these regions.

2 Related works

2.1 RGB-D semantic segmentation

With the widespread application of depth sensors and depth cameras in the field of depth estimation [19-21], people can obtain the depth information of the scene more conveniently, and the research on the image is no longer limited to a single RGB image. The RGB-D semantic segmentation task is to effectively integrate RGB features and depth features to improve segmentation accuracy, especially in some indoor scenes. He et al. [4] proposed an early fusion approach, which simply concatenates an image's RGB and depth channels as a four-channel input to the convolutional neural network. Gupta et al. [6] separately input RGB features and HHA features into two CNNs for prediction and perform fusion in the final stage of the network, and Hazirbas et al. [22] introduced an encoding-decoding network, employing a dual-branch RGB encoder to extract features separately from RGB images and depth images. The studies mentioned above employed equal-weight concatenation or summation operations to fuse RGB and depth features without fully leveraging the complementary information between different modalities. In recent years, some research has proposed more effective strategies for RGB-D feature fusion. Hu et al. [23] utilized a three-branch encoder that includes RGB, Depth, and Fusion branches, efficiently collecting features without breaking the original RGB and deep inference branches. Seichter et al. [24] have presented an efficient RGB-D segmentation approach, characterised by two enhanced ResNet-based encoders utilising an attention-based fusion for incorporating depth information. Fu et al. [25] proposed a joint learning module that learns simultaneously from RGB and depth maps through a shared network, enhancing the model's generalization ability. Fu et al. [25] proposed a joint learning module that learns simultaneously from RGB and depth maps through a shared network, enhancing the model's generalization ability. Zhang et al. [26] proposed a multi-task shared tube structure that aggregates multi-task features into the decoder, improving the learning results for each task. Chen et al. [27] proposed the S-Conv operator, which introduces spatial information to guide the weights of the convolution kernel, thereby adjusting the receptive field, enhancing geometric perception capabilities, and improving segmentation results. Our MIPANet implements a dual-branch convolutional network that performs distinct operations in the middle and final layers of the network to fully utilize the complementary information of different modalities.

2.2 Attention mechanism

In recent years, attention [28-34] has been widely used in computer vision and other fields. Vaswani et al. [28] proposed the self-attention mechanism, which has had a profound impact on the design of the deep learning model. Fu et al. [30] proposed DANet, which can adaptively integrate local features and their global dependencies. Wang et al. [35] utilized spatial attention in an image classification model. Through the backpropagation of a convolutional neural network, they adaptively learned spatial attention masks, allowing the model to focus on the significant regions of the image. Hu et al. [36] proposed channel attention, which adaptively learns the importance of each feature channel through a neural network. Woo et al. [33] incorporated two attention modules that concurrently capture channel-wise and spatial relationships. Wang et al. [37] introduced a straightforward and efficient "local" channel attention mechanism to minimize computational overhead. Qiao et al. [38] introduced a multi-frequency domain attention module to capture information across different frequency domains. Similarly, Ding et al. [39] proposed a contrastive attention module designed to amplify local saliency. Building upon this foundation, Huang et al. [40] proposed a cross-attention module that consolidates contextual information both horizontally and vertically, which can gather contextual information from all pixels. These methods have demonstrated significant potential in single-modality feature extraction. To effectively leverage the complementary information between different modalities, this paper introduces a pooling attention module that learns the differential information between two distinct modalities and fully exploits the intermediate-level features in the convolutional network and the semantic dependencies between modalities.

2.3 Cross-modal interaction

With the development of sensor technology, different types of sensors can provide a variety of modal information for semantic segmentation tasks. The information interaction between RGB and other modalities can improve the performance of multimodal tasks [21, 41–48]. Specifically, Li et al. [21, 41, 42], and Xiao et al. [44] improved the quality of infrared and visible image fusion through cross-modal interaction between RGB image and infrared image. Xiang et al. [45] used a single-shot polarization sensor to build the first RGB-P dataset, incorporated polarization sensing to obtain supplementary information, and improved the accuracy of segmentation for many categories, especially those with polarization characteristics, such as glass. Shen et al. [46] proposed a novel pyramid graph network targeting features, which is closely connected behind the backbone network to explore multi-scale spatial structural features. Shen et al. [47] proposed a structure where graphs and transformers interact constantly, enabling close collaboration between global and local features for vehicle re-identification. Zhuang et al. [48] proposed a network consisting of a two-streams (LiDAR stream and camera stream), which extract features from two modes respectively to realize information interaction between RGB and LIDAR modes. In the task of brain tumor image segmentation, Zhu et al. [49] proposed a new architecture that included an improved Swin Transformer semantic segmentation module, an edge detection module, and a feature fusion module. This design effectively merged deep semantic and edge features, leveraging multi-modal information to integrate global spatial data. Furthermore, Zhu et al. [50] introduced the SDV-TUNet, a model that enriched the network's capacity to handle information by utilizing multi-modal MRI data. They also introduced a multi-level edge feature fusion (MEFF) module, emphasizing the importance of edge information at different levels, which significantly enhanced the precision and efficiency of 3D brain tumor segmentation. Liu et al. [51, 52], fused multi-modal magnetic resonance imaging (MRI) using an adversarial learning framework, treating image fusion as an additional regularization method to aid feature learning, effectively integrating multi-modal features to enhance the model's segmentation performance. Therefore, to fully exploit the features of RGB and Depth images, we advocate for information exchange between these two modalities to leverage their complementary information, thereby enhancing the performance of RGB-D semantic segmentation models.

3 Methods

3.1 Overview

Figure 1 depicts the overall structure of the proposed network. The architecture follows an encoder-decoder design, employing skip connections to facilitate information flow between encoding and decoding layers. The encoder comprises a dual-branch convolutional network, which is used to extract RGB features and depth features. We utilize two pre-trained ResNet50 networks as the backbone, which exclude the final global average pooling layer and fully connected layers. Subsequently, a decoder is employed to upsample the features and integrate them, progressively restoring image resolution.

3.2 Network structure

Given a RGB image $I_{RGB} \in \mathbb{R}^{h\times w\times 3}$, and a Depth image $I_{Dep} \in \mathbb{R}^{h\times w\times 1}$, 3 × 3 convolution is used to extract them shallow features $\boldsymbol{F}^0_{RGB}$ and $\boldsymbol{F}^0_{Dep}$, which can be expressed as Eqs 1 and 2:

$$\boldsymbol{F}^0_{RGB} = Conv_{3\times 3}\left(I_{RGB}\right), \tag{1}$$

$$\boldsymbol{F}^0_{Dep} = Conv_{3\times 3}\left(I_{Dep}\right), \tag{2}$$

where $Conv_{3\times 3}$ denotes 3 × 3 convolution.

The network mainly consists of a four-layer encoder-decoder and introduces two designed modules: MIM and PAM. PAM implements different operations on RGB and depth branches, named PAM-R and PAM-D, respectively. PAM-R refers to PAM in the RGB branch, while PAM-D refers to the PAM in the depth branch. Each layer of the encoder is a ResNetLayer. After $\boldsymbol{F}^0_i$ passing through the ResNetLayer, $\boldsymbol{F}^n_i$ is obtained, the nth layer of the encoder can be expressed as Eq. 3:

$$\boldsymbol{F}^n_i = H^n_i\left(\boldsymbol{F}^{n-1}_i\right), \tag{3}$$

where H^n_i $(n = 1, 2, 3, 4)$ represents the nth ResNetLayer, $i \in \{RGB, Dep\}$ denotes the RGB feature or Depth feature. Specifically, the RGB features and depth features of the first three layers in the ResNet encoder are fed into the PAM. PAM enhances features by performing different operations on RGB features and depth features, resulting in $\tilde{\boldsymbol{F}}^n_{RGB}$ and $\tilde{\boldsymbol{F}}^n_{Dep}$, where $n = 1, 2, 3$. Subsequently, the two features are combined by element-wise addition to obtain $\tilde{\boldsymbol{F}}^n_{Con}$, containing rich spatial location information. Furthermore, the final RGB and depth features from the ResNetLayer4 encoder are fed into the MIM to capture complementary information within these two modalities. The output features of the MIM are then fed into the decoder, where each upsampling layer consists of two 3 × 3 convolutional layers. These layers are followed by batch normalization (BN) and ReLU activation, with each upsampling layer doubling the feature spatial dimensions while halving the number of channels.

3.3 Pooling attention module

Within the low-level features extracted by the convolutional neural network, we capture the fundamental attributes of the input image. These low-level features are critical in modelling the image's foundational characteristics. However, they lack semantic information from the deep-level neural network, such as object shapes and categories. At the same time, during the upsampling process in the decoding layer, there is a risk of losing certain semantic information as the image resolution increases. To address this issue, we introduce the Pooling Attention Module

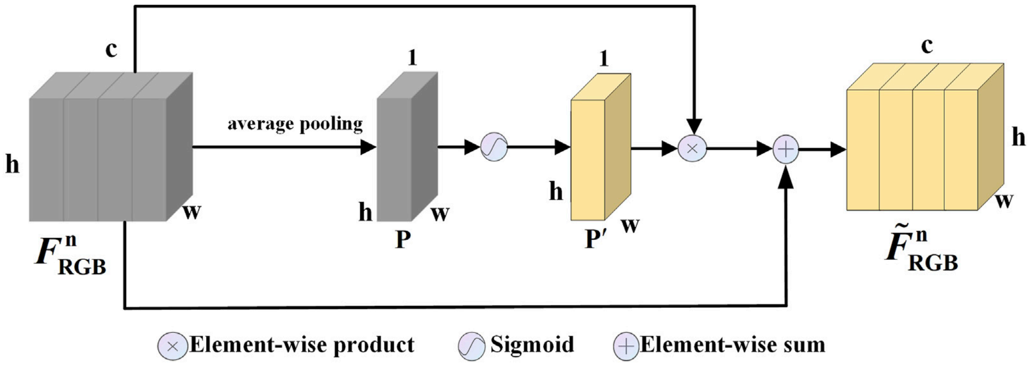


FIGURE 2
Structure of PAM in the RGB branch, referred to as PAM-R. Given an input feature F^n_{RGB}, it is first processed through an average pooling operation to obtain $\boldsymbol{P}$. Subsequently, $\boldsymbol{P}$ undergoes a sigmoid activation to produce $\boldsymbol{P}'$. The activated feature $\boldsymbol{P}'$ is then element-wise product with the original input feature F^n_{RGB} to yield a preliminary result, which is further added to the initial feature F^n_{RGB} to generate the final output $\tilde{F}^n_{RGB}$.

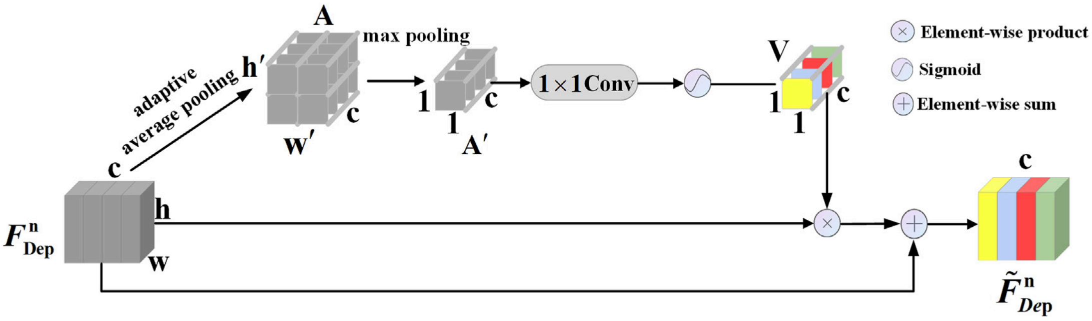


FIGURE 3
Structure of PAM in the depth branch, referred to as PAM-D. The input feature F^n_{Dep} first passes through an adaptive pooling operation, resulting in $\boldsymbol{A}$. This is followed by a max pooling operation to produce $\boldsymbol{A}'$. The output $\boldsymbol{A}'$ then goes through a 1 × 1 convolution and a sigmoid activation to yield the weight vector $\boldsymbol{V}$ (e.g., yellow) between 0 and 1. This $\boldsymbol{V}$ is element-wise product with the original feature F^n_{Dep}, and the product is subsequently added to F^n_{Dep} to produce the final output $\tilde{F}^n_{Dep}$.

(PAM). For RGB features, we utilize average pooling to average the information across all channels at each spatial location. This method highlights the importance of each position, aiding in the better capture of key spatial features such as edges and textures. For depth features, we opt for max pooling, which accentuates the most significant signals within each channel. This effectively enhances the model's response to crucial depth information while suppressing less important channels. This approach allows us to more precisely identify and emphasize important features in the depth map, thus improving the overall segmentation accuracy. In the decoding layer, the output from the PAM is first processed by the Refinement Module (RM), effectively compensating for information loss during the upsampling process, and increasing the network's attention to specific areas. This strategy improves the accuracy of segmentation results and efficiently maintains the integrity of semantic information. The structure of the PAM in RGB and depth branches are shown in Figures 2, 3, respectively.

The input feature $\boldsymbol{F}^n_{RGB} \in \mathbb{R}^{h\times w\times c}$ denotes the RGB feature passes through average pooling to reduce the number of channels in the feature map, which can be expressed as Eq. 4:

$$\boldsymbol{P} = H_{avg}\left(\boldsymbol{F}^n_{RGB}\right), \tag{4}$$

where $\boldsymbol{P} \in \mathbb{R}^{h\times w\times 1}$ represents the feature map that has aggregated the information across all channels at each position. H_{avg} denotes the global average pooling operation. h, w represent the height and width of the feature map. Then we get the weight vector $\boldsymbol{P}' \in \mathbb{R}^{h\times w\times 1}$ by sigmoid activation, which can be expressed as Eq. 5:

$$\boldsymbol{P}' = Sigmoid\left(\boldsymbol{P}\right), \tag{5}$$

Then, we perform an Element-wise product for $\boldsymbol{F}^n_{RGB}$ and $\boldsymbol{P}'$, and the result $\tilde{\boldsymbol{F}}^n_{RGB}$ can be expressed as Eq. 6:

$$\tilde{\boldsymbol{F}}^n_{RGB} = \boldsymbol{F}^n_{RGB} + \left(\boldsymbol{F}^n_{RGB} \otimes \boldsymbol{P}'\right), \tag{6}$$

where ⊗ denotes the Element-wise product. Through the PAM in the RGB branch, the original feature map, after being weighted by spatial attention, is enhanced at important spatial locations, while less important locations are relatively suppressed, thus enabling the network to focus more on spatial regions that are useful for semantic segmentation.

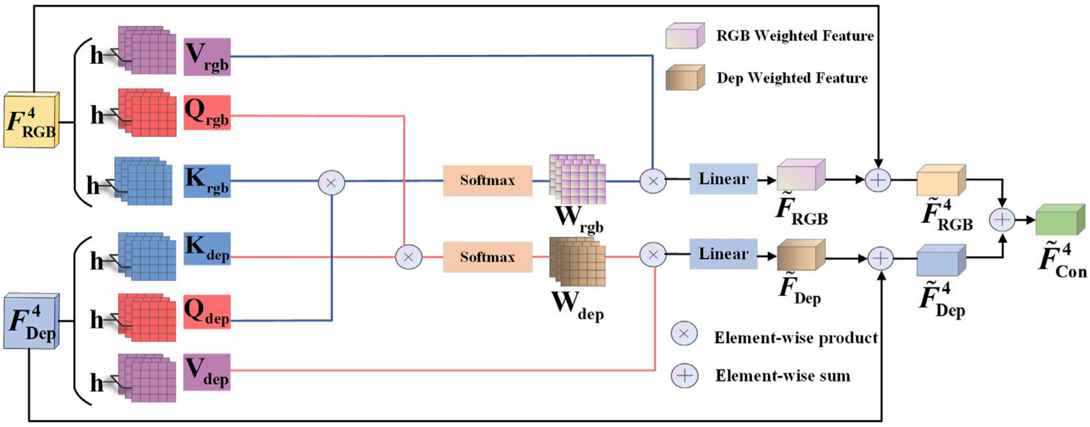


FIGURE 4
Structure of the MIM. The RGB feature and the depth feature undergo linear transformations to generate two sets of Q, K, V (e.g., blue line) for multi-head attention, where h denotes the number of attention heads set to 8. The weighted summation of input features F^4_{RGB} and F^4_{Dep} yields $\tilde{F}^4_{RGB}$ and $\tilde{F}^4_{Dep}$, which are then element-wise added to obtain the output result $\tilde{F}^4_{Con}$.

The input feature $F^n_{Dep} \in \mathbb{R}^{h \times w \times c}$ denotes the Depth feature passes through adaptive average pooling to reduce the feature map to a smaller dimension, which can be expressed as Eq. 7:

$$A = H_{ada}\left(F^n_{Dep}\right), \tag{7}$$

where $A \in \mathbb{R}^{h' \times w' \times c}$ represents the feature map that has been resized by adaptive averaging pooling, H_{ada} denotes the adaptive average pooling operation. h', w' represent the height and width of the output feature map, which we set $h' = 2$ and $w' = 2$. Then, we get the output features A' by max pooling the features after dimensionality reduction, which can be expressed as Eq. 8:

$$A' = H_{max}(A), \tag{8}$$

where $A' \in \mathbb{R}^{1 \times 1 \times c}$ represents the pooling result and then A' undergoes a 1 × 1 convolution and then activation with the sigmoid function, getting a weight vector $V \in \mathbb{R}^{1 \times 1 \times c}$ value between 0 and 1. $H_{\max}$ denotes the max pooling operation. Finally, we perform an Element-wise product for F^n_{Dep} and V, and the result $\tilde{F}^n_{Dep}$ can be expressed as Eqs 9, 10:

$$V = Sigmoid\left(\Phi\left(A'\right)\right), \tag{9}$$

$$\tilde{F}^n_{Dep} = F^n_{Dep} + \left(F^n_{Dep} \otimes V\right), \tag{10}$$

where ⊗ denotes the Element-wise product, Φ denotes 1 × 1 convolution. The PAM in the depth branch makes the network pay more attention to local regions in the image, such as objects near the background in the scene. Meanwhile, adaptive average pooling can enhance the module's flexibility, accommodating diverse input feature map dimensions and fully retaining spatial position information in depth features. $\tilde{F}^n_{Con}$ in Figure 1 can be expressed as Eq. 11:

$$\tilde{F}^n_{Con} = \tilde{F}^n_{RGB} + \tilde{F}^n_{Dep}, \tag{11}$$

During the upsampling process, $\tilde{F}^n_{Con}$ (n = 1, 2, 3) is fed into the decoder.

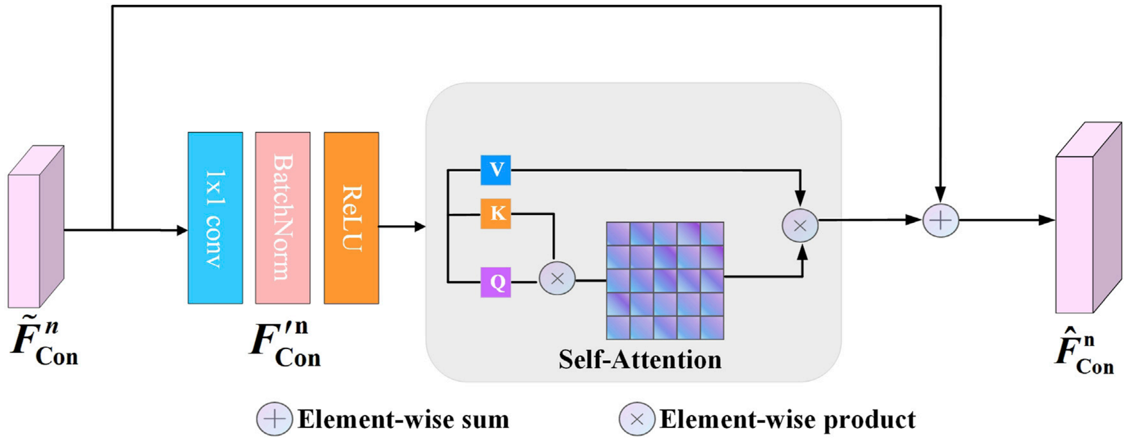
3.4 Multi-modal interaction module

When adjacent objects in an image share similar appearances, distinguishing their categories becomes challenging. Factors such as lighting variations and object overlap, especially in the corners, can lead to their blending with the background. This complexity makes it difficult to precisely identify object edges, leading to misclassification of the object as part of the background. Depth information remains unaffected by lighting conditions and can accurately differentiate between objects and the background based on depth values. Therefore, we design the MIM to supplement RGB information with Depth features. Meanwhile, it utilizes RGB features to strengthen the correlation between RGB and depth features. Depth features excel in capturing object contours and edge information, compensating for the spatial depth information that RGB features lack. Conversely, RGB features play a crucial role in compensating for the deficiencies in depth features, particularly in aspects such as color and texture, thereby enriching the information content of depth features.

MIM achieves dual-mode feature fusion, as depicted in Figure 4. Here, $F^4_{RGB} \in \mathbb{R}^{h \times w \times c}$ and $F^4_{Dep} \in \mathbb{R}^{h \times w \times c}$ correspond to the RGB feature and depth feature from the ResNetLayer4. The feature channels are denoted as "c", and their spatial dimensions are $h \times w$. First, the two feature maps are linearly mapped to generate multi-head query(Q), key(K), and value(V) vectors. Here, "rgb" and "dep" represent the RGB and depth features. These linear mappings are accomplished via fully connected layers, where each attention head possesses its unique weight matrix. For each attention head, we calculate the dot product between two sets of Q and K and then normalize the results to a range between 0 and 1 using the softmax function to get the attention maps W_{rgb} and W_{dep}, which can be expressed as Eqs 12, 13:

$$W_{rgb} = Softmax\left(\frac{Q_{rgb}K^T_{dep}}{\sqrt{d}_k}\right) \tag{12}$$

$$W_{dep} = Softmax\left(\frac{Q_{dep}K^T_{rgb}}{\sqrt{d}_k}\right) \tag{13}$$



$\tilde{\boldsymbol{F}}^{n}_{\text{Con}}$
1x1 conv
BatchNorm
ReLU
$\boldsymbol{F}'^{n}_{\text{Con}}$
V
K
Q
Self-Attention
$\hat{\boldsymbol{F}}^{n}_{\text{Con}}$
Element-wise sum
Element-wise product

FIGURE 5
Design of the Refinement Module: following a 1 × 1 convolution, BatchNorm, and ReLU activation function, the feature map $\boldsymbol{F}'^{n}_{Con}$ is produced. Then, utilizing a self-attention operation, the weighted feature $\hat{\boldsymbol{F}}^{n}_{Con}$ is derived, enhancing the module's ability to focus on salient aspects of the input.

where d_k represents the dimensionality of the K vector. Then, we calculate the RGB weighted feature $\tilde{\boldsymbol{F}}_{RGB}$ and the depth weighted feature $\tilde{\boldsymbol{F}}_{Dep}$, and the final output features $\tilde{\boldsymbol{F}}^{4}_{RGB}$ and $\tilde{\boldsymbol{F}}^{4}_{Dep}$ are obtained through a residual connection, which can be expressed as Eqs 14, 15:

$$\tilde{\boldsymbol{F}}_{RGB} = \boldsymbol{W}_{rgb} \otimes \boldsymbol{V}_{rgb} \tag{14}$$

$$\tilde{\boldsymbol{F}}^{4}_{RGB} = \tilde{\boldsymbol{F}}_{RGB} + \boldsymbol{F}^{4}_{RGB} \tag{15}$$

where $\tilde{\boldsymbol{F}}_{RGB}$ represents the RGB weighted feature, $\boldsymbol{V}_{rgb}$ represents the value vector from the RGB feature, multiplying with weight matrix $\boldsymbol{W}_{rgb}$. $\tilde{\boldsymbol{F}}^{4}_{RGB}$ represents the RGB feature after the fusion with depth feature. Likewise, we get the Eqs 16, 17:

$$\tilde{\boldsymbol{F}}_{Dep} = \boldsymbol{W}_{dep} \otimes \boldsymbol{V}_{dep} \tag{16}$$

$$\tilde{\boldsymbol{F}}^{4}_{Dep} = \tilde{\boldsymbol{F}}_{Dep} + \boldsymbol{F}^{4}_{Dep} \tag{17}$$

where $\tilde{\boldsymbol{F}}_{Dep}$ represents the depth weighted feature, $\boldsymbol{V}_{dep}$ represents the value vector from the Depth feature, multiplying with weight matrix $\boldsymbol{W}_{dep}$. $\tilde{\boldsymbol{F}}^{4}_{Dep}$ represents the depth feature after the fusion with RGB feature, ⊗ represents the Element-wise product. Finally, we can obtain the MIM output through Element-wise sum, which can be expressed as Eq. 18:

$$\tilde{\boldsymbol{F}}^{4}_{Con} = \tilde{\boldsymbol{F}}^{4}_{RGB} + \tilde{\boldsymbol{F}}^{4}_{Dep} \tag{18}$$

3.5 Refinement module

RGB features provide rich colour and texture information, while depth features provide spatial and shape information. The fusion of these two types of features can help the network to understand the scene more comprehensively. However, due to the differences between the two modalities, simple addition might introduce some noise, affecting the segmentation results. To address this issue, we propose a Refinement Module (RM) that, through a CBR structure (Convolution; Batch Normalization; ReLU), allows the network to adaptively re-extract and optimize the fused features, filtering out unnecessary information and retaining features that are more useful for semantic segmentation. Moreover, by utilizing self-attention, the global information of the features is enhanced, enabling a better understanding of the global structure of the input features, thereby improving performance. The structure of the RM is shown in Figure 5.

As shown in Figure 5, $\tilde{\boldsymbol{F}}^{n}_{Con}$ is processed by the CBR operation to generate $\boldsymbol{F}'^{n}_{Con}$, which can be expressed as Eq. 19:

$$\boldsymbol{F}'^{n}_{Con} = CBR\left(\tilde{\boldsymbol{F}}^{n}_{Con}\right) \tag{19}$$

where $n = 1, 2, 3$. CBR represents a 1 × 1 convolution followed by Batch Normalization and ReLU activate function. Then, $\boldsymbol{F}'^{n}_{Con}$ is linearly mapped to generate query(Q), key(K), and value(V) vectors. Through a self-attention module, the final output result is generated, which can be expressed as Eq. 20:

$$\hat{\boldsymbol{F}}^{n}_{Con} = Softmax\left(\frac{\boldsymbol{QK}^{T}}{\sqrt{d}_{k}}\right) \otimes \boldsymbol{V} + \tilde{\boldsymbol{F}}^{n}_{Con} \tag{20}$$

RM further extracts and refines the fused features to enhance the feature representation, and $\hat{\boldsymbol{F}}^{n}_{Con}$ is fed into the decoder.

3.6 Loss function

In this paper, the network performs supervised learning on four different levels of decoding features. We employ nearest-neighbor interpolation to reduce the resolution of semantic labels. Additionally, 1 × 1 convolutions and Softmax functions are utilized to compute the classification probability for each pixel within the output features from the four upsample layers, respectively. The loss function $\mathcal{L}_i$ of layer i is the pixel-level cross entropy loss, which can be expressed as Eq. 21:

$$\mathcal{L}_i = -\frac{1}{N_i}\sum_{\forall p,q} Y(p,q)\log\left(Y'(p,q)\right) \tag{21}$$

where N_i denotes the number of pixels in layer i. p, q represent the coordinate positions of each pixel in the image. Specifically, p refers to the row coordinate of the pixel, while q refers to the column coordinate. Y' is the classification probability of the output, and Y is the label category. The final loss function $\mathcal{L}$ of the network is obtained by summing the pixel-level loss functions of the four decoding layers, which can be expressed as Eq. 22:

$$\mathcal{L} = \sum_{i=1}^{4} \mathcal{L}_i \tag{22}$$

By optimizing the above loss function, the network can get the final segmentation result.

4 Experimental results and analysis

4.1 Experimental setup

NYU-Depth V2 dataset [53] and SUN RGB-D dataset [54] are used to evaluate the proposed method. NYU-Depth V2 dataset is a widely used indoor scene understanding dataset for computer vision and deep learning research. It is an aggregation of video sequences from various indoor scenes recorded by RGB-D cameras from the Microsoft Kinect and is an updated version of the NYU-Depth dataset published by Nathan Silberman and Rob Fergus in 2011. It contains 1,449 RGB-D images, depth images, and semantic tags in the indoor environment. The dataset includes different indoor scenes, scene types, and unlabeled frames, and each object can be represented by a class and an instance number. SUN RGB-D dataset contains image samples from multiple scenes, covering various indoor scenes such as offices, bedrooms, and living rooms. It has 37 categories and contains 10,335 RGB-D images with pixel-level annotations, of which 5,285 are used as training images and 5,050 are used as test images. This special dataset is captured by four different sensors: Intel RealSence, Asus Xtion, Kinect v1, and v2. Besides, this densely annotated dataset includes 146,617 2D polygons, 64,595 3D bounding boxes with accurate object orientations, and a 3D room layout as well as an imaged-based scene category.

We evaluate the results using two standard metrics, Pixel Accuracy (Pix. Acc) and Mean Intersection Over Union (mIoU). Pix. Acc refers to pixel accuracy, which is the simplest metric that represents the proportion of correctly labelled pixels to the total number of pixels, which can be expressed as Eq. 23:

$$Pix.Acc = \frac{\sum_{i=0}^{k} p_{ii}}{\sum_{i=0}^{k}\sum_{j=0}^{k} p_{ij}}. \tag{23}$$

where p_{ii} means to predict the correct value, and p_{ij} means to predict i to j. k represents the number of categories. In addition, Intersection over Union (IoU) is a measure of semantic segmentation, where the IoU ratio of a class is the ratio of the IoU of its true labels and predicted values, while mIoU is the average IoU ratio of each class in the dataset, which can be expressed as Eq. 24:

$$mIoU = \frac{1}{k+1}\sum_{i=0}^{k} \frac{p_{ii}}{\sum_{j=0}^{k} p_{ij} + \sum_{j=0}^{k} p_{ji} - p_{ii}}. \tag{24}$$

where p_{ij} represents the predict i as j, and p_{ji} represents the predict j as i, p_{ii} means to predict the correct value, k represents the number of categories.

We implement and train our proposed network using the PyTorch framework. To enhance the diversity of the training data, we apply random scaling and mirroring. Subsequently, all

TABLE 1 MIPANet compared to the state-of-the-art methods on the NYU-Depth V2 dataset.

Method	Backbone	mIoU (%)	Pix.Acc (%)
ESANet	ResNet18	48.2	—
IEMNet	Res34NBt1D	51.3	76.8
SGACNet	2 × Res34NBt1D	49.4	75.6
Z-ACN	ResNet50	50.0	—
DynMM	ResNet50	51.0	—
RDFNet	2 × ResNet50	47.7	74.8
RAFNet	2 × ResNet50	47.5	73.8
SA-Gate	2 × ResNet50	50.4	—
ESANet	2 × ResNet50	50.5	—
RedNet	2 × ResNet50	47.2	—
ACNet	3 × ResNet50	48.3	—
SGNet	ResNet101	49.6	75.6
RDFNet	2 × ResNet101	49.1	75.6
ShapeConv	ResNet101	51.3	76.4
Baseline	2 × ResNet50	47.4	75.1
Ours (MIPANet)	2 × ResNet50	**52.3**	**77.6**

The bold values mean the highest results.

TABLE 2 MIPANet compared to the state-of-the-art methods on the SUN RGB-D dataset.

Method	Backbone	mIoU (%)	Pix.Acc (%)
IEMNet	Res34NBt1D	48.3	81.9
EMSANet	2 × Res34NBt1D	48.5	—
RAFNet	2 × ResNet50	47.2	81.3
ESANet	2 × ResNet50	48.3	—
RedNet	2 × ResNet50	47.8	81.3
ACNet	3 × ResNet50	48.1	—
SGNet	ResNet101	47.1	81.0
CANet	ResNet101	48.3	82.0
RDFNet	ResNet101	48.2	82.3
ShapeConv	ResNet101	48.6	82.2
RDFNet	2 × ResNet152	47.7	81.5
Baseline	2 × ResNet50	45.5	81.1
Ours (MIPANet)	2 × ResNet50	**49.1**	**82.5**

The bold values mean the highest results.

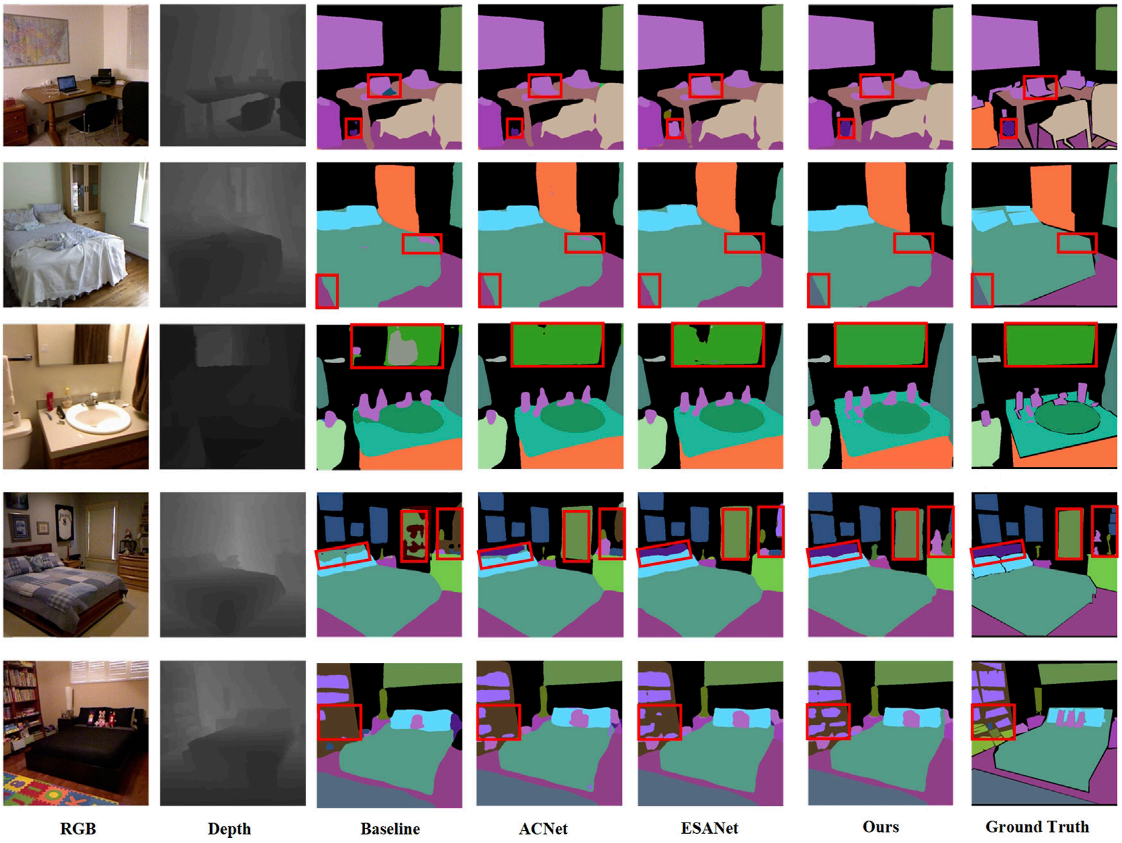


FIGURE 6
Visual comparisons on the NYU-Depth V2 dataset.

RGB and depth images are resized to 480 × 480 for network inputs, and semantic labels are adjusted to sizes of 480 × 480, 240 × 240, 120 × 120, and 60 × 60 for deep supervision training. As the backbone for our encoder, we utilize the ResNet50 pre-trained [55] on the ImageNet dataset [56]. Our baseline model uses two branches as encoders to extract RGB and depth features, respectively, while excluding the PAM during the extraction process. Each branch is composed of four ResNet50 layers. In the final layer of the network, RGB and depth features are merged by element-wise addition, without employing the MIM. The output of element-wise addition is then used as input to the encoder for upsampling operations, resulting in the final segmentation result. To refine the network structure, following [57-59], we adjust it by replacing the 7 × 7 convolution in the input stem with three consecutive 3 × 3 convolutions. The training is conducted on an NVIDIA GeForce GTX 3090 GPU using stochastic gradient descent optimization. Parameters are set with a batch size of 6, an initial learning rate of 0.003, 500 epochs, and momentum and weight decay values of 0.9 and 0.0005, respectively.

4.2 Quantitative experimental results on NYU-Depth V2 and SUN RGB-D datasets

To validate the effectiveness of the proposed model in this paper, we compare the proposed method with state-of-the-arts methods (ESANet [24], IEMNet [60], SGACNet [61], Z-ACN [62], DynMM [63], RDFNet [7], RAFNet [64], SA-Gate [65], RedNet [8], ACNet [23], SGNet [27], ShapeConv [66]) on the NYU-Depth V2 dataset. For a fair comparison, we compare our method with others using the ResNet architecture, which employ ResNet with varying depths and quantities.

Table 1 illustrates our superior performance regarding mIoU and Acc metrics compared to other methods. Specifically, with ResNet50 serving as the encoder in our network, the Pix. Acc and mIoU for semantic segmentation on the NYU-Depth V2 test set reached 77.6% and 52.3%. For example, our method improved the mIoU by 4.9% compared to the baseline method. Compared to the runner-up method DynMMXue and Marculescu (2023), which also employs ResNet50, our method achieved a 1.3% improvement. Similarly, compared to the suboptimal method ShapeConvCao et al. (2021), which uses the deeper ResNet101, our method achieved a 1.0% improvement. Our method achieves better results on networks with ResNet50 as the backbone than some methods with ResNet101 as the backbone, showcasing the effectiveness of our carefully designed network structure.

Then, we compare the proposed method with state-of-the-arts methods (IEMNet [60], EMSANet [67], RAFNet [64], ESANet [24], RedNet [8], ACNet [23], SGNet [27], CANet [68], RDFNet [7], ShapeConv [66]) on the SUN RGB-D dataset. As depicted in Table 2, our approach consistently achieves a higher mIoU score on the SUN RGB-D dataset

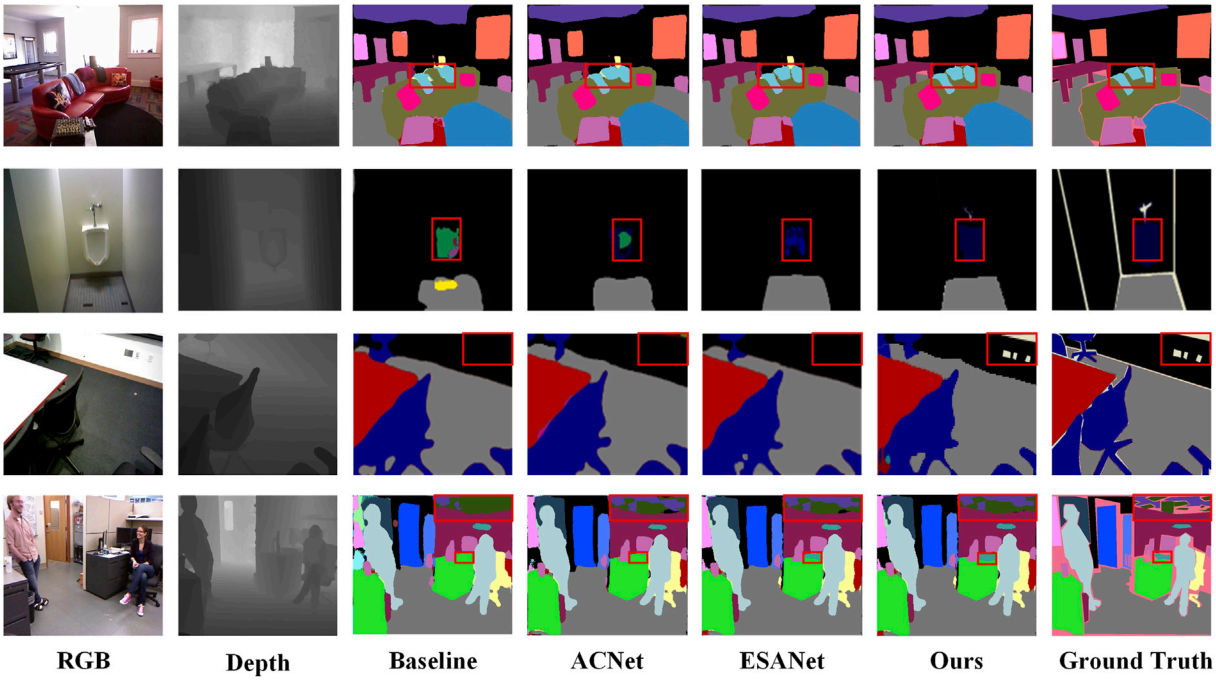


FIGURE 7
Visual comparisons on the SUN RGB-D dataset.

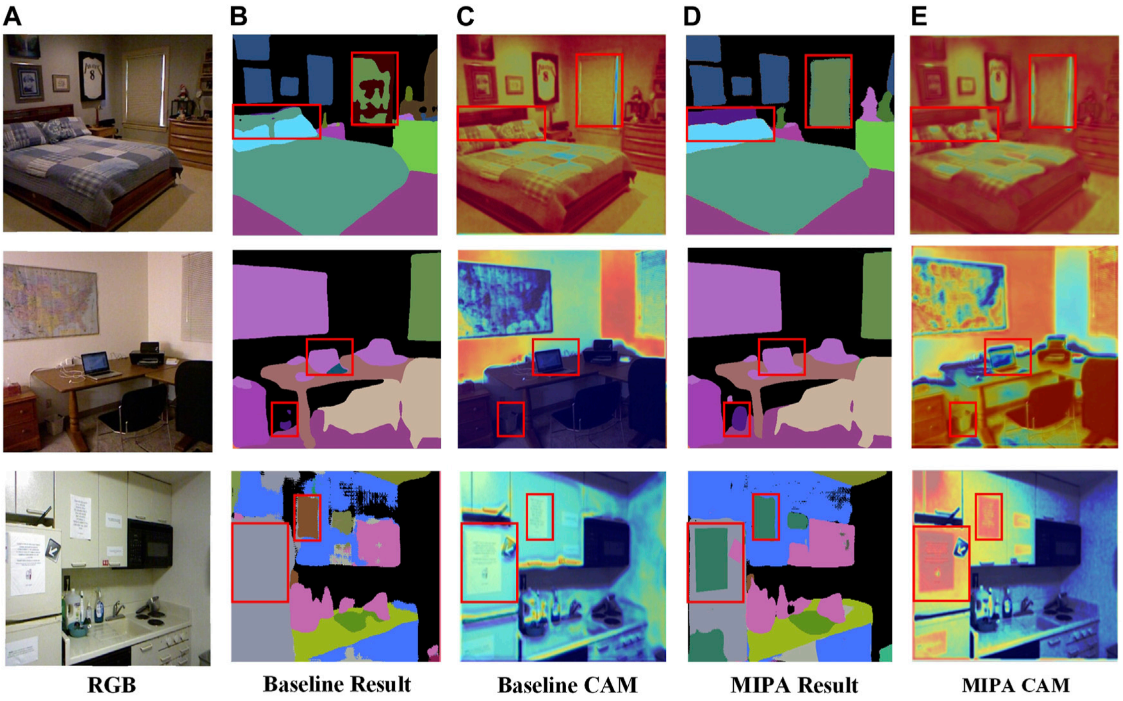


FIGURE 8
Images from left to right represent **(A)** the RGB image, **(B)** the segmentation result of Baseline, **(C)** CAM of Baseline, **(D)** the segmentation results of MIPANet (Ours) and **(E)** CAM of MIPANet. The red box indicates the prominent areas of effect.

TABLE 3 Ablation studies on NYU-Depth V2 dataset for PAM, MIM and RM.

Method	mIoU (%)	Pix.Acc (%)
ResNet50 (Baseline)	47.4	75.1
ResNet50 + PAM	48.9	76.0
ResNet50 + PAM + RM	49.5	76.0
ResNet50 + MIM	51.1	77.0
ResNet50 + PAM + MIM	51.9	77.2
ResNet50 + PAM + MIM + RM (Ours)	52.3	77.6

TABLE 4 Ablation studies on SUN RGB-D dataset for PAM, MIM and RM.

Method	mIoU (%)	Pix.Acc (%)
ResNet50 (Baseline)	45.5	81.1
ResNet50 + PAM	47.9	81.3
ResNet50 + PAM + RM	48.1	81.3
ResNet50 + MIM	48.3	81.5
ResNet50 + PAM + MIM	48.8	82.3
ResNet50 + PAM + MIM + RM (Ours)	49.1	82.5

TABLE 5 Performance comparison of the different methods on the number of model parameters, FLOPs and testing time.

Models	Parameter(M)	FLOPs(G)	Time (ms)
ACNet	116.6	126.3	45.0
RedNet	82.0	101.8	34.7
RDFNet	443.8	648.7	71.9
SA-Gate	110.6	176.5	53.1
MIPANet	360.0	634.2	62.4

than all other methods. For example, our method improved the mIoU by 3.6% compared to the baseline method. Compared to the suboptimal method ESANet [24], which also employs ResNet50, our method achieved a 0.8% improvement. Similarly, compared to the suboptimal method ShapeConv [66], which uses the deeper ResNet101, our method achieved a 0.5% improvement. This observation underscores our module's ability to maintain superior segmentation accuracy, even when dealing with the extensive SUN RGB-D dataset.

4.3 Visualization results on NYU-Depth V2 and SUN RGB-D datasets

To visually highlight the advancements made by our method, we provide visualization results of the network on the NYU-Depth V2 dataset and SUN RGB-D datasets, as shown in Figures 6, 7. From left to right, the RGB image, the Depth image, the baseline model results with ResNet50 backbone, ACNet, ESANet, MIPANet (Ours), and Ground Truth.

As shown in Figure 6, compared to the baseline, our method significantly improve segmentation results. Notably, the dashed box in the figure showcases our network enrich with depth information accurately distinguishes objects from the background. For instance, in the visualization results of the fourth image, the baseline erroneously categorizes the mirror on the wall as part of the background, in the visualization results of the second image, the ACNet and the ESANet mistook the carpet for a part of the floor. In contrast, leveraging depth information, our network discerns the distinct distance information of the mirror from the background, leading to a correct classification of the mirror. The proposed method has achieved precise segmentation outcomes in diverse and intricate indoor scenes. Moreover, it excels in segmenting challenging objects like “carpets” and “books” while delivering finer-edge segmentation results.

As shown in Figure 7, our method also achieve better experimental results on the SUN RGB-D dataset. For example, in the second row of Figure 7, the toilet and wall share a similar white color and partially overlap in position, making it difficult for the network to distinguish between them accurately. Compared to other methods, our MIPA approach demonstrates superior effectiveness in segmenting toilet. In the third row of Figure 7, our method accurately segments the power switch on the wall, further demonstrating its effectiveness.

Furthermore, we verify the effectiveness of our method by providing visualization results of class activation mapping (CAM). These visualizations demonstrate that MIPANet effectively focuses on regions containing adjacent or overlapping objects. As shown in Figure 8, compared to the baseline cam Figure 8C, the more prominent red areas in image Figure 8E indicate that our method focuses more on specific regions. For example, in the first row, the adjacent pillow and headboard are highlighted. In the second row, the trash can overlaps with the wall and has a similar color, the computer is close to the tabletop. In the third row, the paper is attached to the refrigerator and cabinet. The network's attention to these areas increased, compared to the baseline segmentation result in Figure 8B, our method achieves more accurate segmentation results, as shown in Figure 8D. The visualization results indicate that our method better focuses on adjacent and overlapping objects in the image.

4.4 Ablation study

To investigate the impact of different modules on segmentation performance, we conduct ablation experiments on NYU-Depth V2 and SUN-RGBD datasets, as depicted in Tables 3, 4. For instance, in NYU-Depth V2, our PAM module exhibit a superiority of 1.5% and 0.9% over the baseline concerning mIoU and Pix. Acc indicators. Similarly, our MIM module demonstrate a superiority of 3.7% and 1.9% over the baseline regarding mIoU and Pix. Acc. Additionally, the inclusion of the RM has further improved the performance of the module. The result suggests that each proposed module can independently enhance segmentation accuracy. Our module surpasses the baseline in fusing cross-modal features, yielding superior results on both datasets. Using PAM, MIM and RM modules, we achieve the highest mIoU of 52.3% on the NYU-Depth V2 dataset and the highest mIoU of 49.1% on the SUN RGB-D dataset. The result highlights that our designed modules can be collectively optimized to enhance segmentation accuracy.

4.5 Computational complexity analysis

In this section, we analyze the computational complexity of the different methods from three aspects: the number of model parameters, FLOPs, the time required for testing. The results are listed in Table 5. For the evaluation of computational complexity, the size of the input images is standardized to 480 × 640 pixels. The test time is the time taken to process one pair of RGB and depth images. As shown in Table 5, the parameter quantity and FLOPs of our model are moderate. However, compared to the comparison methods, our approach achieves the highest mIoU and exhibits the most visually appealing results.

5 Conclusion

In this paper, we tackle a fundamental challenge in RGB-D semantic segmentation—efficiently fusing features from two distinct modalities. We design an innovative multi-modal interaction and pooling attention network, which uses a small and flexible PAM module in the shallow layer of the network to enhance the feature extraction capability of the network and uses a MIM module in the last layer of the network to integrate RGB features and depth features effectively and then we design a RM during the upsampling stage for feature refinement. The network increases its focus on areas with more potential adjacent objects and overlaps, leading to improvement in the accuracy of RGB-D semantic segmentation. However, due to the attention mechanism adopted by our proposed network, the computational complexity of the network is relatively high. In future research, we will further optimize the network structure to reduce its computational complexity. In addition, we expect to further improve the accuracy of RGB-D segmentation by integrating multiple tasks such as depth estimation and semantic segmentation into a unified framework.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: https://cs.nyu.edu/~fergus/datasets/nyu_depth_v2.html.

Author contributions

SZ: Conceptualization, Formal Analysis, Methodology, Resources, Software, Validation, Visualization, Writing–original draft. MX: Data curation, Investigation, Supervision, Writing–review and editing.

Funding

The authors declare that no financial support was received for the research, authorship, and/or publication of this article.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. Shelhamer E, Long J, Darrell T. Fully convolutional networks for semantic segmentation. *IEEE Trans Pattern Anal Machine Intelligence* (2017) 39:640–51. doi:10.1109/tpami.2016.2572683

2. Li M, Wei M, He X, Shen F. Enhancing part features via contrastive attention module for vehicle re-identification. In: 2022 IEEE International Conference on Image Processing (ICIP); October 16-19, 2022; Bordeaux, France (2022). p. 1816–20.

3. Zhang Z. Microsoft kinect sensor and its effect. *IEEE MultiMedia* (2012) 19:4–10. doi:10.1109/mmul.2012.24

4. He Y, Chiu WC, Keuper M, Fritz M. Std2p: rgbd semantic segmentation using spatio-temporal data-driven pooling. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); July 21 2017 to July 26 2017; Honolulu, HI, USA (2017). p. 7158–67.

5. Couprie C, Farabet C, Najman L, LeCun Y *Indoor semantic segmentation using depth information* (2013). *arXiv preprint arXiv:1301.3572.*

6. Gupta S, Girshick R, Arbeláez P, Malik J. Learning rich features from rgb-d images for object detection and segmentation. *Computer Vision–ECCV 2014: 13th Eur Conf Zurich, Switzerland, September 6-12, 2014, Proc Part VII* (2014) 13:345–60. doi:10.1007/978-3-319-10584-0_23

7. Park SJ, Hong KS, Lee S. Rdfnet: rgb-d multi-level residual feature fusion for indoor semantic segmentation. In: Proceedings of the IEEE international conference on computer vision; 22-29 October 2017; Venice, Italy (2017). p. 4990–9.

8. Lee S, Park SJ, Hong KS. Rdfnet: rgb-d multi-level residual feature fusion for indoor semantic segmentation. In: 2017 IEEE International Conference on Computer Vision (ICCV); 22-29 October 2017; Venice, Italy (2017). p. 4990–9.

9. Eigen D, Fergus R. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In: 2015 IEEE International Conference on Computer Vision (ICCV); 7-13 December 2015; Santiago, Chile (2015). p. 2650–8.

10. Wang A, Lu J, Wang G, Cai J, Cham TJ. Multi-modal unsupervised feature learning for rgb-d scene labeling. In: Computer Vision–ECCV 2014: 13th European Conference; September 6-12, 2014; Zurich, Switzerland (2014). p. 453–67.

11. Shu Z, Li L, Yu J, Zhang D, Yu Z, Wu XJ. Online supervised collective matrix factorization hashing for cross-modal retrieval. *Appl intelligence* (2023) 53:14201–18. doi:10.1007/s10489-022-04189-6

12. Bai Y, Shu Z, Yu J, Yu Z, Wu XJ. Proxy-based graph convolutional hashing for cross-modal retrieval. *IEEE Trans Big Data* (2023) 1–15. doi:10.1109/tbdata.2023.3338951

13. Shu Z, Li B, Mao C, Gao S, Yu Z. Structure-guided feature and cluster contrastive learning for multi-view clustering. *Neurocomputing* (2024) 582:127555. doi:10.1016/j.neucom.2024.127555

14. Li L, Shu Z, Yu Z, Wu XJ. Robust online hashing with label semantic enhancement for cross-modal retrieval. *Pattern Recognition* (2024) 145:109972. doi:10.1016/j.patcog.2023.109972

15. Shu Z, Yong K, Yu J, Gao S, Mao C, Yu Z. Discrete asymmetric zero-shot hashing with application to cross-modal retrieval. *Neurocomputing* (2022) 511:366–79. doi:10.1016/j.neucom.2022.09.037

16. Yang J, Bai L, Sun Y, Tian C, Mao M, Wang G. Pixel difference convolutional network for rgb-d semantic segmentation. *IEEE Trans Circuits Syst Video Tech* (2024) 34:1481–92. doi:10.1109/tcsvt.2023.3296162

17. Zhao Q, Wan Y, Xu J, Fang L. Cross-modal attention fusion network for rgb-d semantic segmentation. *Neurocomputing* (2023) 548:126389. doi:10.1016/j.neucom.2023.126389

18. Yang E, Zhou W, Qian X, Lei J, Yu L. Drnet: dual-stage refinement network with boundary inference for rgb-d semantic segmentation of indoor scenes. *Eng Appl Artif Intelligence* (2023) 125:106729. doi:10.1016/j.engappai.2023.106729

19. Liu F, Shen C, Lin G. Deep convolutional neural fields for depth estimation from a single image. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); June 7 2015 to June 12 2015; Boston, MA, USA (2015). p. 5162–70.

20. Hu J, Huang Z, Shen F, He D, Xian Q. A bag of tricks for fine-grained roof extraction. *IGARSS 2023 - 2023 IEEE Int Geosci Remote Sensing Symp* (2023) 678–80. doi:10.1109/igarss52108.2023.10283210

21. Hu J, Huang Z, Shen F, He D, Xian Q. A rubust method for roof extraction and height estimation. In: IGARSS 2023 - 2023 IEEE International Geoscience and Remote Sensing Symposium; 16 - 21 July, 2023; Pasadena, California, USA (2023). p. 770–1.

22. Hazirbas C, Ma L, Domokos C, Cremers D. Fusenet: incorporating depth into semantic segmentation via fusion-based cnn architecture. *Computer Vis – ACCV* (2017) 2016:213–28. doi:10.1007/978-3-319-54181-5_14

23. Hu X, Yang K, Fei L, Wang K. Acnet: attention based network to exploit complementary features for rgbd semantic segmentation. In: 2019 IEEE International Conference on Image Processing (ICIP); 22-25 September 2019; Taipei, Taiwan (2019). p. 1440–4.

24. Seichter D, Köhler M, Lewandowski B, Wengefeld T, Gross HM. Efficient rgb-d semantic segmentation for indoor scene analysis. In: 2021 IEEE International Conference on Robotics and Automation (ICRA); 30 May - 5 June 2021; Xian, China (2021). p. 13525–31.

25. Fu K, Fan DP, Ji GP, Zhao Q, Shen J, Zhu C. Siamese network for rgb-d salient object detection and beyond. *IEEE Trans Pattern Anal Machine Intelligence* (2022) 44: 5541–59. doi:10.1109/tpami.2021.3073689

26. Zhang X, Zhang S, Cui Z, Li Z, Xie J, Yang J. Tube-embedded transformer for pixel prediction. *IEEE Trans Multimedia* (2023) 25:2503–14. doi:10.1109/tmm.2022.3147664

27. Chen LZ, Lin Z, Wang Z, Yang YL, Cheng MM. Spatial information guided convolution for real-time rgbd semantic segmentation. *IEEE Trans Image Process* (2021) 30:2313–24. doi:10.1109/tip.2021.3049332

28. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. *Adv Neural Inf Process Syst* (2017) 30. doi:10.48550/ARXIV.1706.03762

29. Shen F, Wei M, Ren J *Hsgnet: object re-identification with hierarchical similarity graph network* (2022). arXiv preprint arXiv:2211.05486.

30. Fu J, Liu J, Tian H, Li Y, Bao Y, Fang Z, et al. Dual attention network for scene segmentation. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); June 18 2022 to June 24 2022; New Orleans, LA, USA (2019). p. 3141–9.

31. Shen F, Zhu J, Zhu X, Huang J, Zeng H, Lei Z, et al. An efficient multiresolution network for vehicle reidentification. *IEEE Internet Things J* (2022) 9:9049–59. doi:10.1109/jiot.2021.3119525

32. Shen F, Peng X, Wang L, Hao X, Shu M, Wang Y. Hsgm: a hierarchical similarity graph module for object re-identification. In: 2022 IEEE International Conference on Multimedia and Expo (ICME); July 18 2022 to July 22 2022; Taipei, Taiwan (2022). p. 1–6.

33. Woo S, Park J, Lee JY, Kweon IS. Cbam: convolutional block attention module. In: Proceedings of the European conference on computer vision (ECCV); September 8-14, 2018; Munich, Germany (2018). p. 3–19.

34. Zhang Y, Wang Y, Li H, Li S. Cross-compatible embedding and semantic consistent feature construction for sketch re-identification. In: Proceedings of the 30th ACM International Conference on Multimedia (MM'22); October 10-14, 2022; Lisboa, Portugal (2022). p. 3347–55.

35. Wang F, Jiang M, Qian C, Yang S, Li C, Zhang H, et al. Residual attention network for image classification. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); July 21 2017 to July 26 2017; Honolulu, HI, USA (2017). p. 6450–8.

36. Hu J, Shen L, Sun G. Squeeze-and-excitation networks. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition; June 18 2018 to June 23 2018; Salt Lake City, UT, USA (2018). p. 7132–41.

37. Wang Q, Wu B, Zhu P, Li P, Zuo W, Hu Q. Eca-net: efficient channel attention for deep convolutional neural networks. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); June 13 2020 to June 19 2020; Seattle, WA, USA (2020). p. 11531–9.

38. Qiao C, Shen F, Wang X, Wang R, Cao F, Zhao S, et al. A novel multi-frequency coordinated module for sar ship detection. In: 2022 IEEE 34th International Conference on Tools with Artificial Intelligence (ICTAI); Oct. 31 2022 to Nov. 2 2022; Macao, China (2022). p. 804–11.

39. Ding M, Wang Z, Sun J, Shi J, Luo P. Camnet: coarse-to-fine retrieval for camera re-localization. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV); Oct. 27 2019 to Nov. 2 2019; Seoul, Korea (2019). p. 2871–80.

40. Huang Z, Wang X, Wei Y, Huang L, Shi H, Liu W, et al. Ccnet: criss-cross attention for semantic segmentation. *IEEE Trans Pattern Anal Machine Intelligence* (2023) 45:6896–908. doi:10.1109/tpami.2020.3007032

41. Li H, Cen Y, Liu Y, Chen X, Yu Z. Different input resolutions and arbitrary output resolution: a meta learning-based deep framework for infrared and visible image fusion. *IEEE Trans Image Process* (2021) 30:4070–83. doi:10.1109/tip.2021.3069339

42. Li H, Liu J, Zhang Y, Liu Y. A deep learning framework for infrared and visible image fusion without strict registration. *Int J Comp Vis* (2023) 132:1625–44. doi:10.1007/s11263-023-01948-x

43. Li H, Zhao J, Li J, Yu Z, Lu G. Feature dynamic alignment and refinement for infrared–visible image fusion:translation robust fusion. *Inf Fusion* (2023) 95:26–41. doi:10.1016/j.inffus.2023.02.011

44. Xiao W, Zhang Y, Wang H, Li F, Jin H. Heterogeneous knowledge distillation for simultaneous infrared-visible image fusion and super-resolution. *IEEE Trans Instrumentation Meas* (2022) 71:1–15. doi:10.1109/tim.2022.3149101

45. Xiang K, Yang K, Wang K. Polarization-driven semantic segmentation via efficient attention-bridged fusion. *Opt Express* (2021) 29:4802–20. doi:10.1364/oe.416130

46. Shen F, Zhu J, Zhu X, Xie Y, Huang J. Exploring spatial significance via hybrid pyramidal graph network for vehicle re-identification. *IEEE Trans Intell Transportation Syst* (2022) 23:8793–804. doi:10.1109/tits.2021.3086142

47. Shen F, Xie Y, Zhu J, Zhu X, Zeng H. Git: graph interactive transformer for vehicle re-identification. *IEEE Trans Image Process* (2023) 32:1039–51. doi:10.1109/tip.2023.3238642

48. Zhuang Z, Li R, Jia K, Wang Q, Li Y, Tan M. Perception-aware multi-sensor fusion for 3d lidar semantic segmentation. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV); Oct. 11 2021 to Oct. 17 2021; Montreal, BC, Canada (2021). p. 16260–70.

49. Zhu Z, He X, Qi G, Li Y, Cong B, Liu Y. Brain tumor segmentation based on the fusion of deep semantics and edge information in multimodal mri. *Inf Fusion* (2023) 91: 376–87. doi:10.1016/j.inffus.2022.10.022

50. Zhu Z, Sun M, Qi G, Li Y, Gao X, Liu Y. Sparse dynamic volume transunet with multi-level edge fusion for brain tumor segmentation. *Comput Biol Med* (2024) 172: 108284. doi:10.1016/j.compbiomed.2024.108284

51. Liu Y, Shi Y, Mu F, Cheng J, Chen X. Glioma segmentation-oriented multi-modal mr image fusion with adversarial learning. *IEEE/CAA J Automatica Sinica* (2022) 9: 1528–31. doi:10.1109/jas.2022.105770

52. Liu Y, Mu F, Shi Y, Chen X. Sf-net: a multi-task model for brain tumor segmentation in multimodal mri via image fusion. *IEEE Signal Process. Lett* (2022) 29:1799–803. doi:10.1109/lsp.2022.3198594

53. Silberman N, Hoiem D, Kohli P, Fergus R. Indoor segmentation and support inference from rgbd images. In: Computer Vision–ECCV 2012: 12th European Conference on Computer Vision; October 7-13, 2012; Florence, Italy (2012). p. 746–60.

54. Song S, Lichtenberg SP, Xiao J. Sun rgb-d: a rgb-d scene understanding benchmark suite. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); June 7 2015 to June 12 2015; Boston, MA, USA (2015). p. 567–76.

55. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); June 27 2016 to June 30 2016; Las Vegas, NV, USA (2016). p. 770–8.

56. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, et al. Imagenet large scale visual recognition challenge. *Int J Comput Vis* (2015) 115:211–52. doi:10.1007/s11263-015-0816-y

57. Fu X, Shen F, Du X, Li Z. Bag of tricks for "vision meet alage" object detection challenge. In: 2022 6th International Conference on Universal Village (UV); October 19-22, 2024; Boston, USA (2022). p. 1–4.

58. Shen F, He X, Wei M, Xie Y *A competitive method to vipriors object detection challenge* (2021). arXiv preprint arXiv:2104.09059.

59. Shen F, Wang Z, Wang Z, Fu X, Chen J, Du X, et al. *A competitive method for dog nose-print re-identification* (2022). arXiv preprint arXiv:2205.15934.

60. Xu X, Liu J, Liu H. Interactive efficient multi-task network for rgb-d semantic segmentation. *Electronics* (2023) 12:3943. doi:10.3390/electronics12183943

61. Zhang Y, Xiong C, Liu J, Ye X, Sun G. Spatial information-guided adaptive context-aware network for efficient rgb-d semantic segmentation. *IEEE Sensors J* (2023) 23:23512–21. doi:10.1109/jsen.2023.3304637

62. Wu Z, Allibert G, Stolz C, Ma C, Demonceaux C *Depth-adapted cnns for rgb-d semantic segmentation* (2022). arXiv preprint arXiv:2206.03939.

63. Xue Z, Marculescu R. Dynamic multimodal fusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; June 18 2022 to June 24 2022; New Orleans, LA, USA (2023). p. 2575–84.

64. Yan X, Hou S, Karim A, Jia W. Rafnet: rgb-d attention feature fusion network for indoor semantic segmentation. *Displays* (2021) 70:102082. doi:10.1016/j.displa.2021.102082

65. Chen X, Lin KY, Wang J, Wu W, Qian C, Li H, et al. Bi-directional cross-modality feature propagation with separation-and-aggregation gate for rgb-d semantic segmentation. In: European Conference on Computer Vision; 23-28 August; Glasgow, United Kingdom (2020). p. 561–77.

66. Cao J, Leng H, Lischinski D, Cohen-Or D, Tu C, Li Y. Shapeconv: shape-aware convolutional layer for indoor rgb-d semantic segmentation. In: Proceedings of the IEEE/CVF international conference on computer vision; Oct. 11 2021 to Oct. 17 2021; Montreal, BC, Canada (2021). p. 7068–77.

67. Seichter D, Fischedick SB, Köhler M, Groß HM. Efficient multi-task rgb-d scene analysis for indoor environments. In: 2022 International Joint Conference on Neural Networks (IJCNN); 18-23 July 2022; Padua, Italy (2022). p. 1–10.

68. Tang Q, Liu F, Zhang T, Jiang J, Zhang Y. Attention-guided chained context aggregation for semantic segmentation. *Image Vis Comput* (2021) 115:104309. doi:10.1016/j.imavis.2021.104309