Multimedia Systems (2024) 30:272
https://doi.org/10.1007/s00530-024-01459-w

REGULAR PAPER

Triple fusion and feature pyramid decoder for RGB-D semantic segmentation

Bin Ge[1] · Xu Zhu[1] · Zihan Tang[2] · Chenxing Xia[1] · Yiming Lu[1] · Zhuang Chen[1]

Received: 4 December 2023 / Accepted: 19 August 2024

Abstract
Current RGB-D semantic segmentation networks incorporate depth information as an extra modality and merge RGB and depth features using methods such as equal-weighted concatenation or simple fusion strategies. However, these methods hinder the effective utilization of cross-modal information. Aiming at the problem that existing RGB-D semantic segmentation networks fail to fully utilize RGB and depth features, we propose an RGB-D semantic segmentation network, based on triple fusion and feature pyramid decoding, which achieves bidirectional interaction and fusion of RGB and depth features via the proposed three-stage cross-modal fusion module (TCFM). The TCFM proposes utilizing cross-modal cross-attention to intermix the data from two modalities into another modality. It fuses the RGB attributes and depth features proficiently, utilizing the channel-adaptive weighted fusion module. Furthermore, this paper introduces a lightweight feature pyramidal decoder network to fuse the multi-scale parts taken out by the encoder effectively. Experiments on NYU Depth V2 and SUN RGB-D datasets demonstrate that the cross-modal feature fusion network proposed in this study efficiently segments intricate scenes.

Keywords RGB-D semantic segmentation · Cross-modal · Feature fusion · Attention mechanism

1 Introduction

Semantic segmentation is a fundamental task in computer vision, which aims to assign category labels to each pixel in an image, and plays a very important role in many computer vision tasks, such as automated driving [1, 2], scene understanding [3], medical image segmentation [4], and so on.

So far, Convolutional Neural Network (CNN) based RGB semantic segmentation techniques [5–10] have delivered noteworthy results on many large datasets [11–14]. However, RGB images can only capture the photometric appearance features of the projected image space. Under conditions of poor lighting or similar texture and color of the images, the performance of semantic segmentation methods based on RGB images may decrease significantly. Depth features can provide rich supplementary information for local geometric appearance cues and intuitively reflect the geometry of the visible surface of the object. Researchers have begun to introduce depth information to assist RGB semantic segmentation. With the widespread use of 3D sensors such as Kinect and Xtion, obtaining 3D geometric data on objects has become easier. Therefore, the advantages of enhancing and fusing RGB images and depth images are crucial in semantic segmentation tasks.

In recent years, researchers have focused on improving RGB-D semantic segmentation [15–19] for greater effectiveness. This improvement involves incorporating depth images into the segmentation process. Currently, RGB-D based

✉ Xu Zhu
xzhu@aust.edu.cn

Bin Ge
bge@aust.edu.cn

Zihan Tang
ztang244@wisc.edu

Chenxing Xia
cxxia@aust.edu.cn

Yiming Lu
lqingyi9527@163.com

Zhuang Chen
1609302871@qq.com

[1] College of Computer Science and Engineering, Anhui University of Science and Technology, Huainan 232001, China

[2] University of Wisconsin Madison, Madison, WI, USA

Published online: 16 September 2024

semantic segmentation techniques can be categorized into three main types. The first type is called input layer fusion. It involves using a single encoder network to extract features from both RGB and depth images (as shown in Fig. 1a). However, these methods typically employ simple fusion strategies such as concatenation or element-wise operations like summation or multiplication. For example, Cao et al. [20] proposed using a shape-aware convolution method to handle RGB-D features by concatenating them afterward. The second type is feature layer fusion. The method adopts a dual-stream encoder network structure, with separate encoder networks dedicated to extracting features from RGB and depth images respectively (as shown in Fig. 1b). The extracted features are then combined into a unified representation across multiple scales to facilitate semantic prediction. This two-stream encoder network structure provides great flexibility, allowing researchers to redesign the fusion module or even replace depth images with other types of image modalities such as lidar, thermal infrared, events or line vibration skewness. The third type is output layer fusion. This structure uses a dual-stream encoder network to extract RGB and depth features separately, and uses a fusion module to merge the output layers (as shown in Fig. 1c). However, this method often cannot fully utilize the complementary characteristics of RGB and depth images during the feature extraction process, resulting in unsatisfactory utilization of the fused features of the two modalities.

Feature layer fusion methods are widely favored by researchers for designing RGB-D semantic segmentation network architectures because they offer great scalability and superior segmentation performance. However, effectively combining RGB and depth image features poses a significant challenge due to their different generation mechanisms. The inherent modal differences between RGB and depth images, caused by distinct imaging mechanisms, are often overlooked. This oversight can result in insufficient cross-modal interaction and fusion of complementary information, ultimately affecting the quality of semantic segmentation results.

To address this challenge, we propose a novel framework called TFNet for RGB-D semantic segmentation. TFNet takes RGB and depth images as inputs and utilizes a dual-stream encoder network structure built on the Mix Transformer model to efficiently extract features from both modalities. To ensure the effective fusion of RGB and depth features, we introduce a Three-stage Cross-modal Fusion Module (TCFM). This module facilitates interaction and fusion between RGB and depth features, overcoming the limitations of previous methods that focused solely on one modality. For effective interaction between RGB features and depth features, we design a three-stream self-attention mechanism to achieve effective communication between RGB and depth features. We found that channel selection plays a crucial role in class differentiation, so we introduced a channel-adaptive weighting module to collaborate RGB and depth features based on their relevance. Additionally, drawing inspiration from Xie et al. [21], we incorporate a Feature Pyramid Decoder (FP Decoder) into our framework. This decoder uses the pyramid structure to fully utilize features from all layers of the encoder network, improving segmentation performance by effectively aggregating features from different scales. We evaluate our method on the NYU Depth V2 and SUN RGB-D datasets, demonstrating its effectiveness in segmenting complex scenes with high accuracy and detail.

The main contributions of this paper are as follows.

- We propose an RGB-D semantic segmentation framework, TFNet, to implement RGB-D semantic segmentation by designing an effective fusion module that considers different imaging mechanisms for different modal images.
- We propose a three-stage cross-modal feature fusion module (TCFM). In the first stage, feature interaction is achieved through a cross-modal cross-self-attention mechanism. In the second stage, to promote effective fusion of deep features, an adaptive block is utilized to select feature weights for RGB features. In the final stage, the feature enhancement block utilizes a pyramid network to extract multi-scale features to enhance the fused RGB and depth features.
- We design a lightweight Feature Pyramid Decoder (FP Decoder), which fully utilizes the features extracted in each layer of the encoder network through cascading,

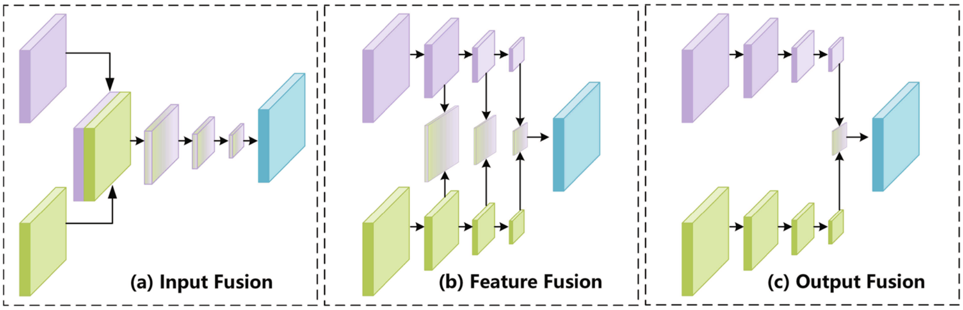


Fig. 1 Comparison of different fusion methods. **a** input layer fusion; **b** feature layer multi-scale fusion; **c** output layer fusion

and effectively aggregating features from different levels.

2 Related work

2.1 RGB semantic segmentation

Traditional semantic segmentation methods take RGB images as input and segment different objects based on pixel perspective. Each pixel of the input RGB image is annotated and predicted into a predefined category. In recent years, deep learning-based models [18, 22–24] for semantic segmentation have become popular and have made significant advancements. These models generally rely on Fully Convolutional Neural Networks [5] (FCNs), which constitute one of the earliest semantic segmentation frameworks that accomplish dense semantic segmentation tasks through an end-to-end pixel-level classification approach. In addition, these models based on Fully Convolutional Networks (FCNs) mainly use pyramid structures such as the Pyramid Pooling Model [25] (PPM) and Atrous Spatial Pyramid Model [7] (ASPP) to capture discriminative multi-scale contextual information from the input images. Although these multi-scale modules for extracting contextual information have been successful in semantic segmentation, they are currently restricted in their receptive domains and cannot effectively extract global semantic information. Noh et al. [26] proposed the first Encoder-Decoder Network (EDN) architecture for semantic segmentation, which is simple and effective for semantic segmentation tasks, and is so far the most popular architecture for semantic segmentation tasks, and is currently used by many state-of-the-art methods. Chaurasia et al. [27] introduced a LinkNet network featuring jump connections within the Encoder-Decoder architecture. The jump connections greatly increase the model speed with minimal accuracy loss, achieving enhanced real-time performance for semantic segmentation tasks. Badrinarayanan et al. [16] designed a new decoder network to improve the performance of semantic segmentation by recovering the low-resolution upsampling into high-resolution feature maps, allowing the network to produce finer segmentation results. There are studies [28–33] that use an encoder-decoder model to integrate multiscale analysis in semantic segmentation networks. Although semantic segmentation based on RGB images has achieved good results, there are great challenges for RGB-based semantic segmentation methods in some conditions where the lighting is poor or the texture and color of the objects are similar, so most researchers nowadays use RGB-D images for semantic segmentation tasks.

2.2 RGB-D semantic segmentation

The depth map corresponding to the RGB image can provide comprehensive geometric and spatial layout information for the RGB image. This can enhance the segmentation performance of complex scenes significantly. Early studies [34, 35] indicate that incorporating depth information can enhance the results of semantic segmentation. Nevertheless, the fusion of depth information into RGB semantic segmentation poses a challenge as the imaging mechanisms of RGB images and depth images differ. Efficiently resolving this challenge remains a matter of inquiry. Some initial methods [36, 37] directly connect the depth image to the RGB image to create a four-channel input for training purposes. Cao et al. [20] previously merged depth and RGB images by concatenation alone. While a shape-based convolution (ShapeConv) was introduced in the network instead of the typical convolution, a single network was insufficient to accurately accommodate the significant discrepancies between the modalities.

To optimally utilize the RGB and depth information, researchers extended the single-stream network structure to a two-stream structure. They used RGB and depth images as inputs to a single network and utilized each stream individually to extract and fuse modality-specific features. This included color and texture information from the RGB image and geometric position information from the depth image. Hazirbas et al. [34] proposed an encoder-decoder based semantic segmentation method, FuseNet, which uses two network branches to simultaneously extract features from both the RGB image and the depth image and improves the performance of semantic segmentation by superimposing the fusion of different levels of RGB features and depth features. Hu et al. [38] have proposed a network framework, named ACNet, consisting of three encoders. The features extracted from RGB images and depth images are merged and passed through the third encoder network. An attention module has also been included. Gupta et al. [39] proposed a method to represent depth information in terms of horizontal parallax, height above ground, and angle of surface normal vector (HHA), and converted the depth images into HHA three-channel type images. Although this method has led to satisfactory outcomes, the HHA coding approach solely focuses on the interdependent information among the data from different channels, and disregards the individually independent parts within each channel. This results in heightened computational volume and some limitations. Chen et al. [40] proposed a spatial information-guided convolution (S-Conv) that can effectively fuse RGB features and HHA features (three-dimensional spatial information features) to enhance the network's perceptual capability.

Although the approaches based on dual-stream decoders stated above may increase performance to some extent, they fail to fully utilize the complementarity of RGB and depth features. To tackle the challenges mentioned previously, we devised a two-stream network featuring two encoder networks based on Mix Transformer. These were designed to efficiently extract RGB image features and depth image features separately. To address the issue of multimodal feature integration, we suggest a three-stage feature fusion module for optimal fusion of the RGB features, which are extracted by the encoder networks, with the depth features. Further details of this module will be provided in Sect. 3.

3 Methods

In this section, we first introduce our proposed TFNet framework for RGB-D semantic segmentation in Sect. 3.1. The proposed three-stage cross-modal fusion module (TCFM) for cross-modal feature fusion is detailed in Sect. 3.2. Our proposed feature pyramid decoder network for feature resolution recovery and category prediction is detailed in Sect. 3.3.

3.1 Architecture overview

We propose a triple fusion network framework, TFNet, for RGB-D semantic segmentation. The framework of TFNet, shown in Fig. 2a, consists of two parallel encoders (RGB Encoder and Depth Encoder) that extract modal features from the RGB image and the depth image, respectively, and then a semantic decoder that recovers the image resolution and predicts the final segmentation result.

Encoder. A dual-stream encoder network structure is designed, where RGB images and depth images are used as inputs to the RGB stream and depth stream networks, respectively, and we use the Mix Transformer (MiT) encoder trained on ImageNet as our backbone network, which is a very powerful and efficient Transformer backbone network. Given the input RGB image and depth image, the encoder first generates patch (block) features through a patch embedding layer. These patch features are passed through four Transformer blocks to produce feature maps with resolutions of $\frac{1}{4}$, $\frac{1}{8}$, $\frac{1}{16}$, and $\frac{1}{32}$, respectively.

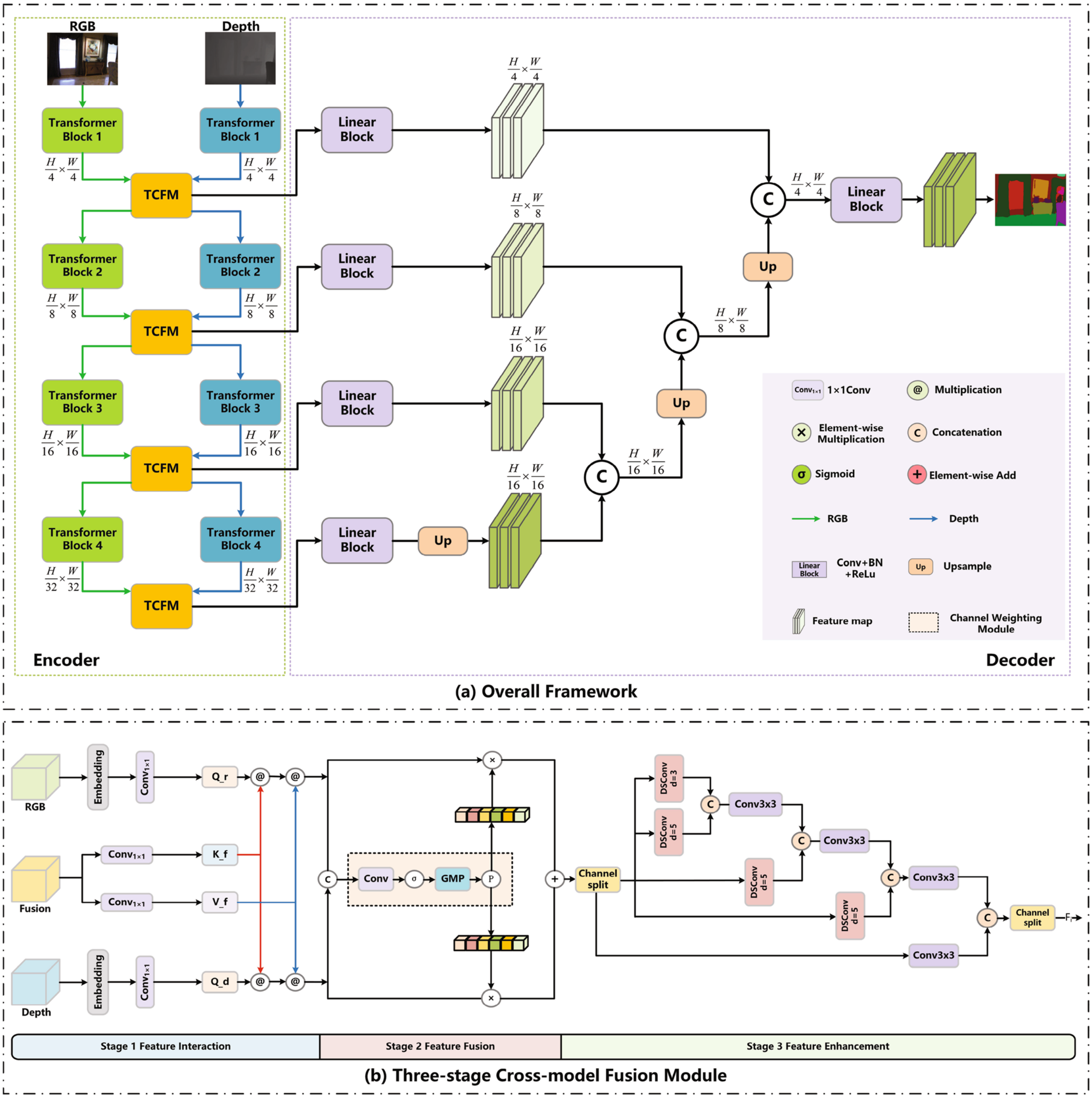
Fusion module. After the output of each Transformer block, it is used to exchange and fuse the information between the extracted features from the RGB encoder and the depth encoder using our proposed three-stage cross-modal fusion module (TCFM). The fusion module takes the inputs of the fusion module from the RGB stream and the depth stream and returns the updated features to the corresponding next Transformer block (as shown in Fig. 2b). More details of our three-stage cross-modal fusion module are described in 3.2.

Decoder.The role of the semantic segmentation decoder is to recover the low-resolution features into high-resolution features to produce the final segmentation result. We designed a Feature Pyramid (FP Decoder) decoder network as the decoder network for the segmentation task, and the simple network design has high efficiency. The specific structural details of the decoder network will be described in detail in Sect. 3.3.

3.2 Three-stage cross-modal fusion module

Because the imaging mechanisms of RGB and depth images are not the same, RGB features and depth features are fundamentally different, with long-range contextual correlation and global spatial consistency for RGB data and local geometric consistency for depth data. Despite the large difference between the two modal features, there is a large amount of complementary information between RGB features and depth features. To effectively fuse the features of two different modalities, we construct a three-stage cross-modal fusion module (TCFM) to effectively interact and fuse the RGB and depth features (shown in Fig. 2b). In the feature interaction phase, RGB and depth still maintain two branches for feature interaction through the cross-modal cross-attention mechanism. In the feature fusion phase, we use channel weighting to weight and fuse RGB features and depth features into one feature. Finally, in the feature enhancement phase, we redesign the Distribution Shifting Convolution (DSConv) structure inspired by Gennari et al. [41] to enhance the fused feature by changing the receptive field of the convolution operation.

Feature interaction (FI). As analyzed above, although the imaging mechanisms of RGB images and depth images are not the same, the semantic information from different modalities is usually complementary. Usually, we can interact the semantic information of one modality with that of the other modality to achieve feature enhancement in each modality. However, previous attention-based approaches were only performed based on separate modalities and did not take full advantage of the fusion features of the two modalities. Therefore, this paper proposes a novel and effective cross-modal cross-self-attention module using three features (RGB, Depth, and RGB+Depth) for interaction, which can fully realize the information interaction between RGB and Depth modalities. In the feature interaction stage, the features of the two modalities (RGB features and depth features) interact through a symmetric dual-path structure. Specifically, this study integrates RGB and depth features into the fusion feature F_{fused}, creating an additional branch. Four independent convolutional layers are used in this paper to generate self-aware Queries, Keys and Values from RGB, depth, and fused features, respectively. Subsequently, the



(a) Overall Framework

(b) Three-stage Cross-model Fusion Module

Fig. 2 **a** Overall framework of TFNet. The inputs are RGB images and depth images. **b** Detailed architecture of the three-stage cross-modal fusion module (TCFM)

keys and values generated from the fused feature are multiplied with the RGB features and depth features to produce interacted RGB features and depth features. For easier comprehension, only the RGB modal path is illustrated, as shown in Eq. (1).

$$\begin{aligned} Q_r &= \mathcal{N}(\mathcal{R}(Conv(F_{RGB}))),\\ K_f &= \mathcal{N}(\mathcal{R}(Conv(F_{fused}))),\\ V_f &= \mathcal{R}(Conv(F_{fused})),\\ F_{RGB} &= Q_r @ K_f^T @ V_f, \end{aligned} \tag{1}$$

Where F_{RGB} denotes the RGB feature, F_{fused} denotes the fusion feature, $\mathcal{N}(\cdot)$ denotes normalisation, $\mathcal{R}(\cdot)$ denotes

Rearrange, $Conv(\cdot)$ denotes convolution, and @ denotes matrix multiplication.

Feature fusion (FF). The simplest fusion strategies based on RGB-D semantic segmentation are element summation and concatenation, but these methods cannot effectively exploit the complementary features of RGB and Depth. Some researchers have used more complex fusion strategies, such as obtaining the fusion features of RGB and Depth by weighted summation, but the weighting in these studies is to set the same weights for all channels. These weights have a good promoting effect on some very high-quality depth maps, but the effect on some low-quality depth maps is not good and even plays an inhibiting effect. In this context, we design a fusion module based on channel weighting, which re-evaluates the importance of different channels by reweighting them based on the channels, and effectively fuses RGB features and depth features.

The RGB and depth features are concatenated in the first stage, followed by weighting using a channel-adaptable weighting block to assign weights to the RGB channels and depth feature channels. This channel-weighting block accurately determines the significance of each RGB and depth channel. The specific operation is shown in Eq. (2).

$$\begin{aligned} W_r &= GMP(\sigma(Conv(\mathcal{C}(F_{RGB}, F_{Depth})))), \\ \hat{fused}^i &= W_r * F_{RGB} + (1 - W_r) * F_{Depth}, \end{aligned} \tag{2}$$

where F_{RGB} denotes the RGB feature, F_{Depth} denotes the depth feature, $GMP(\cdot)$ denotes global mean pooling, $\sigma(\cdot)$ denotes the Sigmoid activation function, $Conv(\cdot)$ denotes convolution, and $\mathcal{C}(\cdot)$ denotes the concatenate operation, and $*$ denotes element-wise multiplication.

Feature enhancement (FE). Inspired by channel shuffling, we perform further feature enhancement on the fused features, using channel separation and channel shuffling, and finally feed them into a four-branch feature enhancement module. The feature enhancement module is defined as shown in Eq. (Results per class for3):

$$fused^i = FE(\hat{fused}^i) \tag{3}$$

where $\hat{fused}^i$ denotes the fused features generated in the feature fusion stage, $FE(\cdot)$ is the feature enhancement module, which is an enhancement of the receptive field block (RFB), as shown in stage 3 in Fig. 2b. We first split the fused features into two parts in the dimension of the channel by a channel-splitting operation. We feed one part to a pyramid network to extract its multi-scale features, effectively enhancing the extracted features. The other part is connected to the features obtained in the first part by residuals using a 1 × 1 convolution. Finally, we use a channel shuffle operation to ensure information communication between different channels.

3.3 Feature pyramid decoder network

Using the feature pyramid network structure, we have designed a feature pyramid decoder (FP Decoder). The feature pyramid network has the characteristic of having different resolutions at different scales, and targets of different sizes can have appropriate feature representations at the corresponding scales, and then by fusing the multi-scale information, targets of different sizes can be analyzed. The feature pyramid decoder uses multiple Linear blocks and upsampling to restore the features $F_i(i = 1, 2, 3, 4)$ extracted at each stage of the encoder to the $\frac{H}{4} \times \frac{W}{4} \times C$ size, and then passes to get the predicted image F.

The proposed feature pyramid decoder consists of four main steps (the structure of the feature pyramid decoder is shown in Fig. 2a). We unify the multilayer features from the encoder network to unify the dimensions through multiple linear layers, and then generate the predicted image through a mask prediction. The designed decoder can be expressed as:

$$\begin{aligned} \hat{F}_4 &= Up(\mathcal{L}(F_4)), \\ \hat{F}_3 &= Up(\mathcal{C}(\hat{F}_4, \mathcal{L}(F_3))), \\ \hat{F}_2 &= Up(\mathcal{C}(\hat{F}_3, \mathcal{L}(F_2))), \\ F &= \mathcal{L}(\mathcal{C}(\hat{F}_2, \mathcal{L}(F_1))), \end{aligned} \tag{4}$$

where $F_i(i = 1, 2, 3, 4)$ is the fusion features generated by the TCFM module, F is the predicted prediction map, $\mathcal{L}(\cdot)$ is the linear layer, $Up(\cdot)$ is the upsampling operation, and $\mathcal{C}(\cdot)$ denotes the concatenate operation.

4 Experiments

In this section, we present experimental results to verify the effectiveness of our proposed TFNet method for RGB-D semantic segmentation. In Sect. 4.1, two publicly available RGB-D semantic segmentation datasets, NYU Depth V2 [42] and SUN RGB-D [43], are briefly introduced and the two main evaluation metrics for RGB-D semantic segmentation are briefly described. Section 4.2 describes some details of the experiments. In Sect. 4.3, the performance of our proposed RGB-D semantic segmentation task model TFNet on the RGB-D datasets NYU Depth V2 and SUN RGB-D is demonstrated and compared with state-of-the-art methods. Section 4.4 verifies the validity of our proposed three-stage cross-modal fusion module and decoder network. Additionally, we provide a range of qualitative results that serve to enhance the analysis of segmentation results.

4.1 Dataset and metrics

We assess the efficacy of our proposed network through training and evaluation of two widely used indoor RGB-D semantic segmentation datasets.

NYU Depth V2 dataset. The NYU Depth V2 dataset comprises video images depicting a variety of indoor scenes obtained through RGB and Depth cameras from Microsoft Kinect. Raw depth images are captured via Microsoft Kinect sensor, while optimized depth images are generated using the proposed colouring scheme in the publication. The dataset incorporates 1449 RGB-D images with a resolution size of 640 × 480, split into a training set of 795 samples and a test set of 654 samples. The semantic categories are primarily divided into 13 and 40 categories, and we use the most common to annotate the 40 semantic categories in most current studies.

SUN RGB-D dataset. The SUN RGB-D dataset is a comprehensive resource for RGB-D scene understanding tasks. It comprises newly captured data and integrates samples from various existing datasets, such as NYU Depth V2 [42], Berkeley B3DO [44], and SUN3D [45]. The dataset comprises 10,335 indoor RGB-D images, organized into a training set and a test set with 5285 and 5050 samples respectively. All images have densely annotated 37 semantic labels.

We employ two widely used metrics for result evaluation: Pixel Accuracy (pixel Acc.) and Mean Intersection Over Union (mIoU). Pixel Accuracy carries significant weight in semantic segmentation tasks.

Pixel Accuracy represents the proportion of accurately segmented pixels to the total number of pixels in the semantic segmentation image. This is calculated using Eq. (5).

$$PixelAcc. = \frac{\sum_{i=0}^{n} p_{ii}}{\sum_{i=0}^{n} \sum_{j=0}^{n} p_{ij}} \quad (5)$$

The mean Intersection over Union (mIoU) measures the extent of overlap between the segmentation outcome and the actual image by averaging the ratio between the intersection and union of predicted and ground-truth pixel regions. This metric frequently assesses the success of semantic image segmentation, with Eq. (6) depicting the precise calculations for mIoU.

$$mIoU = \frac{1}{n+1} \sum_{i=0}^{n} \frac{p_{ii}}{\sum_{j=0}^{n} p_{ij} + \sum_{j=0}^{n} p_{ji} - p_{ii}} \quad (6)$$

4.2 Implementation details

We have trained and implemented our network utilizing the Pytorch framework. For the encoder, we utilize the default configuration of Mix Transformer. Throughout the training, we employed AdamW as our optimizer and a poly learning rate with a coefficient of 0.9 and an initial learning rate of $6e^{-5}$. Below, we specify the particularities of different datasets.

NYU Depth V2 dataset. The model was trained on NVIDIA V100 GPU using MiT-B2 backbone, with a training epoch number of 500. The entire image of 640 × 480 size was used for both training and inference. A batch size of 8 was applied for the MiT-B2 backbone.

SUN-RGBD dataset. The model was trained on a NVIDIA V100 GPU. The training epoch number was set to 200. The image was randomly cropped to 480 × 480. A batch size of 8 was used for the MiT-B2 backbone.

4.3 Comparison with state-of-the-arts

We compare our proposed framework with existing state-of-the-art methods on two public datasets.

(1) Comparison results on NYUv2: As shown in Table 1, TFNet achieves the highest mIoU of 53.6%. Our experimental results, measured by two segmentation metrics (mIoU and Pixel Acc.), are comparable to those of more advanced networks currently available. Most of the previous methods for feature fusion are not aware of the differences in the imaging mechanisms of RGB and depth images, and cannot effectively utilise RGB and depth features. The results in Table 1 show that the TCFM module in TFNet is reasonable and efficient. Furthermore, we can obtain more crucial

Table 1 Performance comparison on the NYU Depth V2 (Class 40) dataset

Method	mIoU (%)	Pixel Acc. (%)
3DGNN [46]	43.1	–
LS-DeconvNet [47]	45.9	71.9
CFN [48]	47.7	–
ACNet [41]	48.3	–
SGNet [40]	51.1	76.8
ShapeConv [20]	51.3	76.8
NANet [49]	52.3	77.9
SA-Gate [50]	52.4	77.9
DCANet [51]	53.3	78.2
CMANet [52]	47.6	74.2
ConvNeXt-CMFFM [53]	51.9	76.8
SGACNet [54]	49.4	75.6
Ni et al. [55]	51.1	77.3
TFNet (our)	53.6	78.2

indicators for evaluating the model by focusing on the similarities and differences between the two modalities.

Table 2 presents the results of our comparison of the accuracy of each type of mIoU on the NYU Depth V2 dataset. Our focus is not only on accuracy, but also on the distribution of data for each category. Surprisingly, we have to admit that the results of our method in some categories (e.g. contain, lamp, picture) are satisfactory. Our TFNet benefits from TCFM, which extracts useful feature information from RGB images and depth images to obtain effective

Table 2 Results per class for RGB-D semantic segmentation on the NYU Depth V2 dataset

Class	Method						
	RDFNet [56]	CANet [17]	SGACNet [54]	Model-1	Model-2	Model-3	TFNet(our)
Wall	79.7	79.6	80.3	80.0	**81.4**	*83.1*	*83.1*
Floor	87.0	87.5	87.7	85.0	*88.7*	**88.6**	*88.7*
Cabinet	60.9	61.1	60.9	62.3	60.6	*65.3*	**64.3**
Bed	73.4	70.7	71.5	68.4	71.9	72.6	*73.6*
Chair	64.6	63.7	64.6	60.1	64.4	**64.8**	*65.6*
Sofa	*65.4*	**64.7**	63.6	58.0	62.9	64.2	64.0
Table	50.7	46.8	45.8	47.0	49.3	**52.3**	*52.4*
Door	39.9	*44.6*	39.9	39.7	39.8	**44.3**	42.1
Window	49.6	46.5	49.0	49.7	49.4	*52.2*	**51.7**
Bookshelf	44.9	46.9	44.9	**47.7**	44.3	47.5	*47.8*
Picture	61.2	61.2	61.9	63.1	64.8	*65.9*	**65.8**
Counter	67.1	68.9	67.1	65.0	67.2	*71.7*	**70.6**
Blinds	63.9	58.0	60.1	61.7	*66.2*	63.6	**64.4**
Desk	*28.6*	22.4	24.9	22.0	20.2	**25.7**	24.7
Shelves	14.2	14.1	**18.8**	18.6	17.4	*19.5*	17.6
Curtain	59.7	56.1	60.6	63.2	**68.9**	65.3	*69.6*
Dresser	49.0	47.0	51.2	47.1	40.7	**53.6**	*56.1*
Pillow	*49.9*	**48.6**	47.9	41.3	46.4	42.2	45.0
Mirror	54.3	49.1	49.1	40.1	*57.7*	**54.8**	52.0
Floor mat	*39.4*	32.0	*39.4*	**39.1**	47.2	38.9	38.7
Cloths	24.4	22.9	21.5	24.3	**26.2**	24.9	*27.9*
Ceiling	66.0	79.0	76.4	72.8	78.8	*81.1*	**80.7**
Books	33.0	32.7	32.6	35.2	34.3	**35.5**	*35.9*
Refridgerator	52.4	51.8	52.0	55.4	57.0	**60.5**	*67.7*
Television	52.6	**60.4**	50.3	56.3	54.7	59.5	*66.5*
Paper	31.3	32.7	33.0	34.0	*34.9*	33.2	**34.2**
Towel	36.8	38.4	40.6	41.5	43.1	*49.1*	**46.4**
Shower curtain	23.6	41.3	43.2	27.1	40.3	**43.2**	*49.5*
Box	11.1	**14.7**	11.5	*15.0*	13.5	13.3	14.4
Whiteboard	63.7	*81.9*	57.1	76.2	75.8	69.3	**77.6**
Person	78.6	81.0	76.7	80.5	79.9	**81.7**	*83.0*
Night stand	38.6	39.0	**49.3**	45.2	44.7	43.2	*49.4*
Toilet	68.4	78.0	76.5	74.9	74.4	*81.1*	78.8
Sink	53.2	61.9	**65.7**	57.5	58.5	*68.5*	63.4
Lamp	45.9	49.5	51.4	49.4	51.1	**53.8**	*55.6*
Bathtub	32.9	53.5	54.9	45.2	49.9	*65.1*	**61.2**
Bag	**14.6**	9.3	9.8	8.8	11.3	13.6	*16.5*
Otherstructure	32.9	28.1	31.9	32.5	31.8	**34.0**	*35.0*
Otherfurniture	18.7	20.1	*21.1*	19.2	18.1	**20.5**	20.2
Otherprop	36.4	39.3	39.2	39.5	41.6	**41.8**	*42.2*

The results in the table are percentages of mIoU. The top two results are shown in italics and bold. Where Model-1 denotes the baseline,Model-2 denotes the removal of the feature pyramid decoder only, and Model-3 denotes the removal of the TCFM module only fusion features for segmentation. Finally, our TFNet has the highest mIoU values in 19 of the 40 classes in Table 2. The results show that our method is more reliable in balancing the segmentation performance.

Time complexity and space complexity are the main criteria for evaluating model efficiency. Therefore, we compare TFNet with [40, 52, 54] to verify the efficiency of the model. As shown in Table 3, compared with CMANet, our method outperforms CMANet by 6.0% in terms of performance, reduces parameters by 48.6%, and reduces FLOPs by 69.4%. At the same time, TFNet is better than SGNet. With the parameter amount reduced by 6.3% and FLOPs increased by 39.5%, TFNet's mIoU increased by 2.4% and Pixel Acc. increased by 1.4%. Although our parameters and FLOPs have increased compared with SGACNet, TFNet's mIoU has increased by 4.2%. It can be seen that TFNet has achieved a balance between model complexity and accuracy, and our future research will further improve the balance between complexity and accuracy.

(2) Comparison results on SUN RGB-D: The SUN RGB-D dataset is a substantial dataset with a greater number of training and testing samples than the NYU Depth V2 dataset, making it more demanding. Recent studies show minimal variations in the segmentation outcomes of various methods on the SUN RGB-D dataset. However, our new approach outperforms the previous methods (as illustrated in Table 4), demonstrating its ability to generalise to larger datasets.

4.4 Ablation study

In order to study the functionality of the proposed network and its processing modules, extensive ablation experiments are conducted on the NYU Depth V2. Each experiment used the same hyper-parameter settings during experiments.

Baseline. Our framework employs a Mix Transformer backbone for RGB-D semantic segmentation tasks. In order to show the contribution made by the method proposed in this paper, a baseline was designed for the ablation study (As shown in Model-1 in Table 5). The baseline uses the RGB and depth image as inputs, and each encoder extracts features it simply fuses the RGB feature with the depth feature by simple elemental summation, and uses the fused features as an input to the decoder. In addition, the decoder network up-samples the fused features extracted at each stage and finally uses the join operation (as shown in Fig. 3).

Triple fusion and feature pyramid decoder. We perform ablation experiments on our proposed network architecture TFNet to investigate the impact of our proposed three-stage cross-modal fusion module as well as the feature pyramid decoder on segmentation accuracy. In both training and testing, each experiment is performed with the same set of hyperparameters for ablation experiments. Table 5 shows the performance of each part of our proposal. The experimental results show that using TCFM+FP-Decoder, the result of mIoU is 53.6%, which is an improvement of 4.8% over the baseline. In addition, to verify the effectiveness of each stage in the TCFM module in TFNet, we designed three defective models (Model-4, Model-5, and Model-6). These three models all remove part of the TCFM module. Model-4 removes Stage 2 and Stage 3 in TCFM, Model-5 removes Stage 1 and Stage 3, and Model-6 removes Stage 1. Among them, Model-4, Model-5, and Model-6 improved by 2.9%, 2.8%, and 3.4% respectively compared with the baseline. The results show that each stage in the TCFM module improves the accuracy of semantic segmentation, and the

Table 4 Performance comparison on the SUN RGB-D dataset

Method	mIoU (%)	Pixel Acc. (%)
3D-GNN [49]	45.9	–
D-CNN [57]	42.0	–
ACNet [38]	48.1	–
SGNet [40]	48.5	81.8
NANet [49]	48.8	82.3
ShapeConv [20]	47.6	82.0
EMSANet [58]	48.4	–
CMANet [52]	47.2	81.1
Link-RGBD [59]	48.4	**83.1**
SGACNet [54]	47.8	81.2
Ni et al. [55]	49.0	81.7
TFNet (our)	**49.2**	82.5

The top result is shown in bold

Table 3 Comparison of model complexity

Method	Architecture base	Params/M	FLOPs/G	mIoU (%)	Pixel Acc. (%)
SGNet [40]	ResNet-101	64.7	26.0	51.2	76.8
CMANet [52]	ResNet-50	117.8	137.2	47.6	74.2
SGACNet [54]	ResNet-34	33.5	16.7	49.4	75.6
TFNet(our)	Transformer	60.6	42.6	53.6	78.2

The results are reported in terms of the number of parameters (million), the computing complexity of FLOPs (gigabytes), mean IoU (%), and Pixel Acc. (%)

Fig. 3 Overall framework of TFNet-Baseline (Model-1)

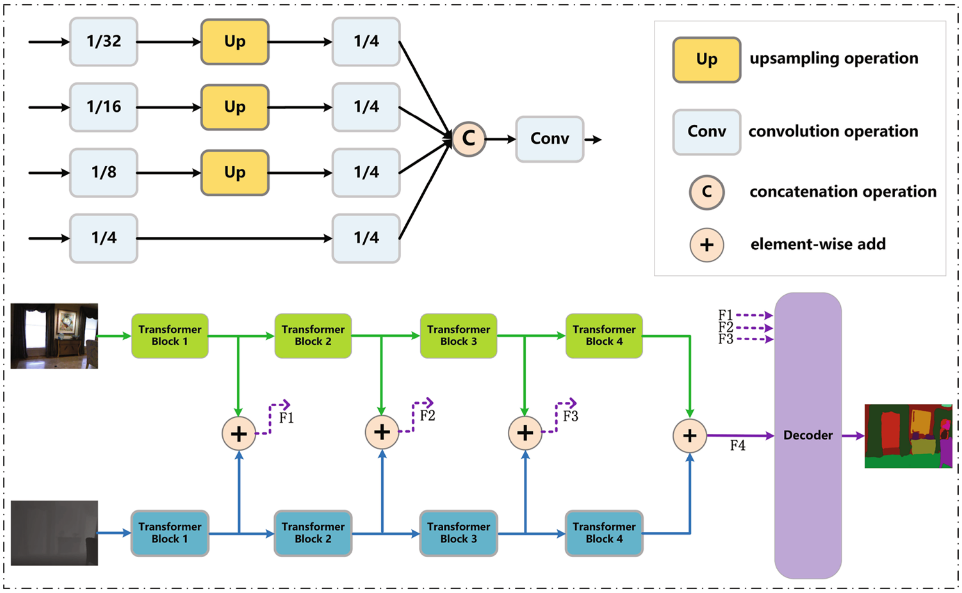


Table 5 TFNet ablation experiments on the NYU Depth V2 dataset

Model	TCFM	FP-Decoder	Pixel Acc	mIoU
Model-1			75.4	48.8
Model-2	✓		78.0	52.7
Model-3		✓	76.8	50.7
Model-4	✓ (Stage1)	✓	77.4	51.7
Model-5	✓ (Stage2)	✓	77.2	51.6
Model-6	✓ (Stage2–3)	✓	77.7	52.2
TFNet (our)	✓	✓	78.2	53.6

Where Model-1 is the baseline

Table 6 Ablation for TCFM on the NYU Depth V2 dataset

Model	Attention mechanism	Pixel Acc. (%)	mIoU (%)
Model-6	CA + SA	77.9	52.8
TFNet (our)	TCFM (stage 1)	78.2	53.6

SA and CA denote spatial attention and channel attention, respectively

original TFNet has the best effect, proving that the proposed TCFM module is effective. In the fusion module, the various stages also promote each other. In addition, in order to verify the effectiveness of the designed feature pyramid decoder, a decoder that directly upsamples the restored size is designed, as shown in Fig. 3 (the upper left). The results show that the FP-Decoder can effectively improve the performance of semantic segmentation.

In addition, in order to verify the effectiveness of the three-stream self-attention mechanism proposed in this paper, the stage 1 in the TCFM is replaced with the current common cross-attention mechanism, and the results show that the attention mechanism proposed in this paper has better results (as shown in Table 6).

Quantitative and qualitative analysis must be performed to more accurately evaluate the performance of semantic segmentation. Therefore, we visualize the experimental results. Therefore, we visualize the experimental results. In Fig. 4, we visualize some classic semantic segmentation examples using our baseline, defective model, and TFNet, aiming to enhance the understanding of our segmentation results. As shown in (a) and (d) in Fig. 4, for places with strong lighting, the baseline and three defect models cannot accurately segment. For the black parts in (a), (b), and (d), classification errors may occur in baseline, Model-4 and Model-5. For some small items in (c), (e), and (f), various defect models cannot effectively segment them. The TFNet we proposed rarely suffers from the above defects. Therefore, the experimental results in Fig. 4 verify the effectiveness of our proposed TFNet.

5 Conclusion

In this paper, we present TFNet, a network framework designed for indoor scene segmentation, specifically aimed at resolving cross-modal fusion challenges. TFNet comprises two primary modules: the three-level cross-modal fusion module (TCFM) and the lightweight feature pyramid decoder network (FP Decoder). These modules are seamlessly integrated into an encoder-decoder network architecture. The three-stage fusion module incorporates a unique three-branch structure with a cross-attention mechanism to facilitate effective interaction between RGB

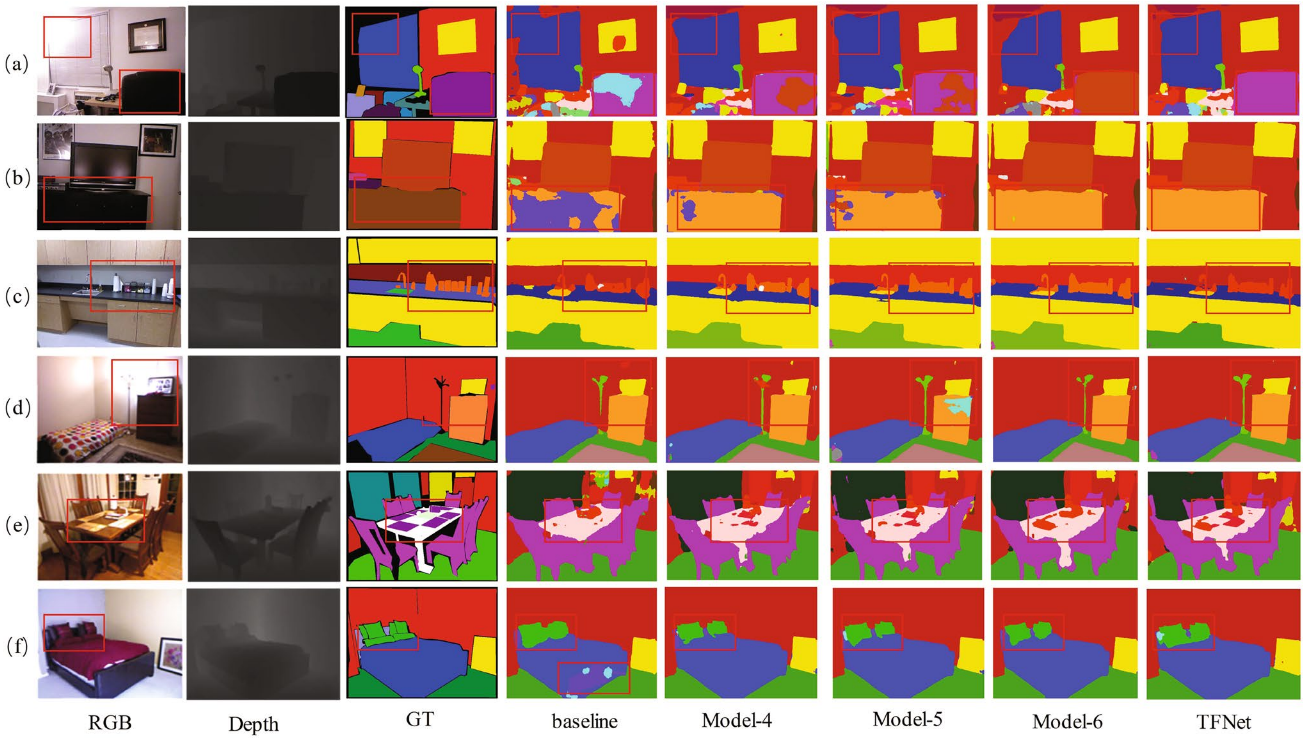


Fig. 4 Visualisation of results on NYU Depth V2 dataset

and depth features. It also employs the weighted fusion of RGB and depth features via the channel adaptive weighting module. Additionally, we propose a feature pyramid decoder-based pyramid network, which adeptly fuses features across multiple scales. Our semantic segmentation method adopts a simplified decoder design to enhance efficiency and effectiveness while maintaining accuracy.

We conduct extensive experiments on various challenging indoor RGB-D datasets to validate the effectiveness of our semantic segmentation method. While there is potential for further enhancements in the accuracy of RGB-D semantic segmentation, we prioritize meeting real-time performance requirements, particularly for applications like autonomous driving. Hence, our future research will concentrate on improving accuracy while ensuring efficient computing speed.

Author contributions Zhu was responsible for writing the manuscript text and experimental design, Ge and Xia for revising the manuscript, Tang for data analysis, Lu for experimental design and data collection, and Chen for organizing the experimental data.

Funding Funding was provided by National Natural Science Foundation of China (Grant No. 62102003).

Data availability No datasets were generated or analysed during the current study.

Declarations

Conflict of interest The authors declare no conflict of interest.

References

1. Yang, L., Liang, X., Wang, T., Xing, E.: Real-to-virtual domain unification for end-to-end autonomous driving. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 530–545 (2018)
2. Xiao, X., Zhao, Y., Zhang, F., Luo, B., Yu, L., Chen, B., Yang, C.: Baseg: boundary aware semantic segmentation for autonomous driving. Neural Netw. **157**, 460–470 (2023)
3. López-Cifuentes, A., Escudero-Vinolo, M., Bescós, J., García-Martín, Á.: Semantic-aware scene recognition. Pattern Recognit. **102**, 107256 (2020)
4. Wei, J., Wu, Z., Wang, L., Bui, T.D., Qu, L., Yap, P.-T., Xia, Y., Li, G., Shen, D.: A cascaded nested network for 3T brain MR image segmentation guided by 7T labeling. Pattern Recognit. **124**, 108420 (2022)
5. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440 (2015)
6. Ronneberger, O., Fischer, P., Brox, T.: U-net: convolutional networks for biomedical image segmentation. In: Medical Image

Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III 18, pp. 234–241. Springer (2015)

7. Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Trans. Pattern Anal. Mach. Intell. **40**(4), 834–848 (2017)
8. Xu, X., Li, G., Xie, G., Ren, J., Xie, X., et al.: Weakly supervised deep semantic segmentation using CNN and ELM with semantic candidate regions. Complexity **2019**, 9180391 (2019)
9. Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2117–2125 (2017)
10. Lin, X., Sánchez-Escobedo, D., Casas, J.R., Pardàs, M.: Depth estimation and semantic segmentation from a single RGB image using a hybrid convolutional neural network. Sensors **19**(8), 1795 (2019)
11. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (VOC) challenge. Int. J. Comput. Vis. **88**, 303–338 (2010)
12. Mottaghi, R., Chen, X., Liu, X., Cho, N.-G., Lee, S.-W., Fidler, S., Urtasun, R., Yuille, A.: The role of context for object detection and semantic segmentation in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 891–898 (2014)
13. Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: common objects in context. In: Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13, pp. 740–755. Springer (2014)
14. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3213–3223 (2016)
15. 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, pp. 1925–1934 (2017)
16. Badrinarayanan, V., Kendall, A., Cipolla, R.: Segnet: a deep convolutional encoder–decoder architecture for image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **39**(12), 2481–2495 (2017)
17. Zhou, H., Qi, L., Huang, H., Yang, X., Wan, Z., Wen, X.: CANet: co-attention network for RGB-D semantic segmentation. Pattern Recognit. **124**, 108468 (2022)
18. Ying, X., Chuah, M.C.: Uctnet: uncertainty-aware cross-modal transformer network for indoor RGB-D semantic segmentation. In: European Conference on Computer Vision, pp. 20–37. Springer (2022)
19. 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. Intell. **125**, 106729 (2023)
20. 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, pp. 7088–7097 (2021)
21. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: Segformer: simple and efficient design for semantic segmentation with transformers. Adv. Neural Inf. Process. Syst. **34**, 12077–12090 (2021)
22. Romeo, L., Devanna, R., Marani, R., Matranga, G., Biddoccu, M., Milella, A.: Scale-invariant semantic segmentation of natural RGB-D images combining decision tree and deep learning models. In: Multimodal Sensing and Artificial Intelligence: Technologies and Applications III, vol. 12621, pp. 257–260. SPIE (2023)
23. Yoon, J., Han, J., Nguyen, T.P.: Logistics box recognition in robotic industrial de-palletising procedure with systematic RGB-D image processing supported by multiple deep learning methods. Eng. Appl. Artif. Intell. **123**, 106311 (2023)
24. Li, Y., Ouyang, S., Zhang, Y.: Combining deep learning and ontology reasoning for remote sensing image semantic segmentation. Knowl.-Based Syst. **243**, 108469 (2022)
25. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2881–2890 (2017)
26. Noh, H., Hong, S., Han, B.: Learning deconvolution network for semantic segmentation. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1520–1528 (2015)
27. Chaurasia, A., Culurciello, E.: Linknet: exploiting encoder representations for efficient semantic segmentation. In: 2017 IEEE Visual Communications and Image Processing (VCIP), pp. 1–4. IEEE (2017)
28. Ding, H., Jiang, X., Shuai, B., Liu, A.Q., Wang, G.: Context contrasted feature and gated multi-scale aggregation for scene segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2393–2402 (2018)
29. He, J., Deng, Z., Qiao, Y.: Dynamic multi-scale filters for semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3562–3572 (2019)
30. He, J., Deng, Z., Zhou, L., Wang, Y., Qiao, Y.: Adaptive pyramid context network for semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7519–7528 (2019)
31. Huang, Z., Wang, C., Wang, X., Liu, W., Wang, J.: Semantic image segmentation by scale-adaptive networks. IEEE Trans. Image Process. **29**, 2066–2077 (2019)
32. Knolle, M., Kaissis, G., Jungmann, F., Ziegelmayer, S., Sasse, D., Makowski, M., Rueckert, D., Braren, R.: Efficient, high-performance semantic segmentation using multi-scale feature extraction. PLoS ONE **16**(8), 0255397 (2021)
33. Li, S., Wan, L., Tang, L., Zhang, Z.: Mfeafn: multi-scale feature enhanced adaptive fusion network for image semantic segmentation. PLoS ONE **17**(9), 0274249 (2022)
34. Hazirbas, C., Ma, L., Domokos, C., Cremers, D.: Fusenet: incorporating depth into semantic segmentation via fusion-based CNN architecture. In: Computer Vision—ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, November 20–24, 2016, Revised Selected Papers, Part I 13, pp. 213–228. Springer (2017)
35. Jiang, J., Zheng, L., Luo, F., Zhang, Z.: Rednet: residual encoder–decoder network for indoor RGB-D semantic segmentation (2018). arXiv preprint arXiv:1806.01054
36. Eigen, D., Fergus, R.: Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2650–2658 (2015)
37. 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, pp. 4837–4846 (2017)
38. 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), pp. 1440–1444. IEEE (2019)
39. Gupta, S., Girshick, R., Arbeláez, P., Malik, J.: Learning rich features from RGB-D images for object detection and segmentation. In: Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part VII 13, pp. 345–360. Springer (2014)

40. Chen, L.-Z., Lin, Z., Wang, Z., Yang, Y.-L., Cheng, M.-M.: Spatial information guided convolution for real-time RGBD semantic segmentation. IEEE Trans. Image Process. **30**, 2313–2324 (2021)
41. Nascimento, M.G.d., Fawcett, R., Prisacariu, V.A.: Dsconv: efficient convolution operator. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 5148–5157 (2019)
42. 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, Florence, Italy, October 7–13, 2012, Proceedings, Part V 12, pp. 746–760. Springer (2012)
43. 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, pp. 567–576 (2015)
44. Janoch, A., Darrell, T., Abbeel, P., Malik, J.: The berkeley 3d object dataset. Techn. Report No. UCB/EECS-2012-85, University of California at Berkeley (2012)
45. 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, pp. 1625–1632 (2013)
46. 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, pp. 5199–5208 (2017)
47. Wang, J., Wang, Z., Tao, D., See, S., Wang, G.: Learning common and specific features for RGB-D semantic segmentation with deconvolutional networks. In: Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part V 14, pp. 664–679. Springer (2016)
48. Lin, D., Chen, G., Cohen-Or, D., Heng, P.-A., Huang, H.: Cascaded feature network for semantic segmentation of RGB-D images. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1311–1319 (2017)
49. Zhang, G., Xue, J.-H., Xie, P., Yang, S., Wang, G.: Non-local aggregation for RGB-D semantic segmentation. IEEE Signal Process. Lett. **28**, 658–662 (2021)
50. Yu, L., Gao, Y., Zhou, J., Zhang, J., Wu, Q.: Multi-layer feature aggregation for deep scene parsing models (2020). arXiv preprint arXiv:2011.02572
51. Bai, L., Yang, J., Tian, C., Sun, Y., Mao, M., Xu, Y., Xu, W.: Dcanet: differential convolution attention network for RGB-D semantic segmentation (2022). arXiv preprint arXiv:2210.06747
52. 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 **22**(21), 8520 (2022)
53. Tang, X., Li, B., Guo, J., Chen, W., Zhang, D., Huang, F.: A cross-modal feature fusion model based on convnext for RGB-D semantic segmentation. Mathematics **11**(8), 1828 (2023)
54. Zhang, Y., Xiong, C., Liu, J., Ye, X., Sun, G.: Spatial-information guided adaptive context-aware network for efficient RGB-D semantic segmentation. IEEE Sens. J. **23**, 23512–23521 (2023)
55. Ni, J., Zhang, Z., Shen, K., Tang, G., Yang, S.X.: An improved deep network-based RGB-D semantic segmentation method for indoor scenes. Int. J. Mach. Learn. Cybern. **15**, 589–604 (2023)
56. 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, pp. 4980–4989 (2017)
57. Wang, W., Neumann, U.: Depth-aware CNN for RGB-D segmentation. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 135–150 (2018)
58. Seichter, D., Fischedick, S.B., Köhler, M., Groß, H.-M.: Efficient multi-task RGB-D scene analysis for indoor environments. In: 2022 International Joint Conference on Neural Networks (IJCNN), pp. 1–10. IEEE (2022)
59. Wu, P., Guo, R., Tong, X., Su, S., Zuo, Z., Sun, B., Wei, J.: Link-RGBD: Cross-guided feature fusion network for RGBD semantic segmentation. IEEE Sens. J. **22**(24), 24161–24175 (2022)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.