Image and Vision Computing 142 (2024) 104892

Contents lists available at ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

EFDCNet: Encoding fusion and decoding correction network for RGB-D indoor semantic segmentation

Jianlin Chen [a,b], Gongyang Li [a,b,*], Zhijiang Zhang [a,b,*], Dan Zeng [a,b]

[a] *Shanghai Institute for Advanced Communication and Data Science, Shanghai University, Shanghai 200444, China*
[b] *School of Communication and Information Engineering, Shanghai University, Shanghai 200444, China*

ARTICLE INFO

Keywords:
RGB-D indoor semantic segmentation
Encoding fusion
Decoding correction

ABSTRACT

Semantic segmentation is a crucial task in vision measurement systems that involves understanding and segmenting different objects and regions within an image. Over the years, numerous RGB-D semantic segmentation methods have been developed, leveraging the encoder-decoder architecture to achieve outstanding performance. However, existing methods have two main problems that constrain further performance improvement. Firstly, in the encoding stage, existing methods have a weak ability to fuse cross-modal information, and low-quality depth maps can easily lead to poor feature representation. Secondly, in the decoding stage, the upsampling of high-level semantic information may cause the loss of contextual information, and low-level features from the encoder may bring noises to the decoder through skip connections. To solve these issues, we propose a novel *Encoding Fusion and Decoding Correction Network* (EFDCNet) for RGB-D indoor semantic segmentation. First, in the encoding stage of EFDCNet, we focus on extracting valuable information from low-quality depth maps, and employ a channel-wise filter to select informative depth features. Additionally, we establish the global dependencies between RGB and depth features via the self-attention mechanism to enhance the cross-modal feature interactions, extracting discriminant and powerful features. Then, in the decoding stage of EFDCNet, we use the highest-level information as semantic guidance to compensate for the upsampling information and filter out noise from the low-level encoder features propagated through the skip connections to the decoder. Extensive experiments conducted on two widely-used RGB-D indoor semantic segmentation datasets demonstrate that the proposed EFDCNet surpasses the performance of relevant state-of-the-art methods. The code is available at https://github.com/ Mark9010/EFDCNet

1. Introduction

Semantic segmentation is a critical task in vision measurement systems that plays a crucial role in enabling machines to analyze and comprehend visual content at a fine-grained level [1]. By segmenting images at the pixel level and associating them with semantic labels, semantic segmentation provides a detailed understanding of objects, regions, and their relationships within an image [2–4]. It has great practical utility in many applications such as autonomous driving [5], inspection robotics [6], and virtual/augmented reality [7]. Many effective pure RGB semantic segmentation solutions have been proposed in the past decade. However, they cannot handle scenes with the influence of lighting changes, occlusion, and unclear object edges well. RGB-D sensor technology represented by Kinect can provide accurate position and depth information of object surfaces [8], which has played an important role in indoor inspection robot navigation as shown in Fig. 1. The depth image captured by Kinect can help solve the above challenging scenes [9,10]. Therefore, RGB-D indoor semantic segmentation [11] is gaining attention from researchers because of its increasing importance and its potential for significant impact in various applications, including robotics, augmented reality, smart environments, and more.

For RGB-D semantic segmentation, the design of the encoder can be divided into three types: early fusion [12,13], late fusion [14–16], and multi-level fusion [17–19], depending on the order in which the two modalities are fused. Early fusion combines the two modalities before the encoder. Late fusion extracts features from both modalities, and fuses them at high levels. Multi-level fusion uses a dual-stream structure in the encoder stage to extract RGB and depth features separately and fuses them at every level. Currently, multi-level fusion strategy has

[*] Corresponding authors at: Shanghai Institute for Advanced Communication and Data Science, Shanghai University, Shanghai 200444, China.
E-mail addresses: chen1026@shu.edu.cn (J. Chen), ligongyang@shu.edu.cn (G. Li), zjzhang@staff.shu.edu.cn (Z. Zhang), dzeng@shu.edu.cn (D. Zeng).

https://doi.org/10.1016/j.imavis.2023.104892
Received 25 August 2023; Received in revised form 15 November 2023; Accepted 20 December 2023
Available online 26 December 2023

shown better performance and is a research focus in this field.

In the multi-level fusion methods, researchers focus on how to effectively integrate cross-modal RGB and depth features in the encoder stage. One approach treats the depth map as auxiliary information and utilizes the positional information of the depth map to adjust the RGB features [20–22]. However, in this approach, the interaction and correlation between RGB and depth information are insufficiently modeled, making the encoder have a weak ability to leverage the complementary nature of these two modalities. Another approach equally treats cross-modal RGB and depth information, and fuses the RGB and depth feature in a symmetrical structure [17–19,23]. This approach ignores the low-quality characteristics of the depth map, which may introduce noise into the cross-modal fusion. These two approaches result in the inability to extract valuable features in the encoder stage, restricting performance improvement.

In the field of semantic segmentation, the prevalent design for the decoding stage involves the skip connection and the upsampling operator [24,25]. That is, the features from the encoder will be connected to the features of the corresponding level of the decoder via the skip connection. This manner allows for the preservation of detailed information, which is beneficial for the pixel-level classification task, i.e., the semantic segmentation task. The upsampling operator is used to gradually restore the resolution of features in a level-by-level manner until the original size is reached. However, most current methods [24,26] only utilize addition or concatenation to connect the features of encoder and decoder. Such simple operations may introduce noise from the low-level encoder features into the decoding stage. Moreover, the upsampling operator usually directly adopts interpolation or deconvolution without the assistance of other information [27,28], which can lead to loss of texture and detail information [29], as well as confusion of object positions.

Based on the above analysis, in this paper, we attempt to address the aforementioned issues from two aspects: encoding fusion and decoding correction. Firstly, in the encoding stage, we design an encoding fusion module to enhance the capability of cross-modal feature extraction. This module employs channel-wise filters [30] to depth features, with the goal of suppressing irrelevant or noisy clues originating from the depth map and preserving informative clues. Additionally, based on the self-attention mechanism [31], we establish global dependencies between RGB and depth features. This enables the encoder to facilitate cross-modal feature interactions, generating discriminative fused features. Secondly, in the decoding stage, we design a decoding correction module to achieve feature correction. This module utilizes high-level semantic information to correct the features by compensating for the decoder upsampling features and filtering out the noise from the encoder skip-connecttion features. Based on the above two modules, we propose a novel and effective solution, named Encoding Fusion and Decoding Correction Network (EFDCNet), for RGB-D indoor semantic segmentation, which can effectively alleviate the above issues and achieve promising performance.

Our main contributions are summarized as follows:

- We propose a novel *Encoding Fusion and Decoding Correction Network* (EFDCNet) for RGB-D indoor semantic segmentation. Our EFDCNet performs feature fusion in the encoding stage and feature correction in the decoding stage to achieve multi-modal information fusion and decoding feature correction, generating high-quality semantic segmentation maps.
- We propose an *Encoding Fusion Module* (EFM) for the encoder. EFM can extract valuable information from depth maps and establish global dependencies between RGB and depth features, enhancing cross-modal interactions and generating discriminant features.
- We propose a *Decoding Correction Module* (DCM) for the decoder. DCM corrects features in the decoder with the semantic information from two aspects: compensating for information loss during feature upsampling and filtering out the noises in the encoder features.

2. Related work

2.1. RGB semantic segmentation

Long et al. [1] first proposed a fully convolutional network for semantic segmentation. Subsequently, the U-shaped encoder-decoder architecture [24] sparked further research interest. In this architecture, the encoder captures information at different scales and abstraction levels, while the decoder is responsible for gradually upsampling and reconstructing the output segmentation map to the same resolution as the input image. To improve network stability and prevent gradient disappearance or explosion, both U-Net [24] and SegNet [25] used skip connections in the symmetrical structure to fuse encoding information into the decoder. Later, a series of improvement works emerged, such as multiscale aggregation, global context encoding, and more.

Effectively utilizing multiscale information can significantly enhance the receptive field. Chen et al. [32] proposed the atrous spatial pyramid pooling (ASPP) to aggregate global multiscale information. Tsai et al. [33] designed a short dense connection network as the backbone to gradually downsample and re-aggregate multiscale feature maps. Global context encoding can effectively improve segmentation performance. Zhao et al. [34] introduced a hierarchical pyramid pooling module to address the issue of context information loss between different subregions. Zhang et al. [35] proposed a context encoding module to capture global contextual cues and emphasize category-specific information relevant to the scene. In addition, low-level local contextual features from shallow layers are also important for identifying small objects and distinguishing boundaries. Zhang et al. [29] focused on combining different levels of information to enhance feature fusion. Fu et al. [36] attached more local context to positions with lower similarity to global features and repeatedly used gate-controlled local features. Li et al. [37] used attention guidance for global enhancement and local refinement, enhancing the context information.

Overall, RGB semantic segmentation methods focused on exploring the specific designs in encoder-decoder structures. However, for complex scenes with severe lighting changes and occlusions, the performance will be greatly limited [38]. Therefore, it would be meaningful to

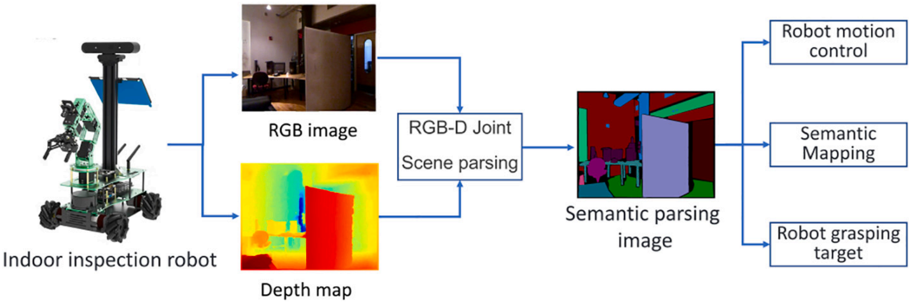

Fig. 1. The typical workflow diagram for a commercial indoor inspection robot.

introduce additional depth information and study joint semantic segmentation of RGB-D data.

2.2. RGB-D semantic segmentation

RGB-D data plays a crucial role in various computer vision applications, including 3D Reconstruction [39], action Recognition [40], human tracking [41] and salient object detection [42–45]. Depth maps contain valuable and comprehensive geometric distance information that is useful for dealing with low contrast and cluttered background scenes. Therefore, joint segmentation has received much attention. In RGB-D semantic segmentation, the key is where and how to fuse them.

Existing RGB-D semantic segmentation methods can be classified into early fusion, late fusion, and multi-level fusion according to the location of cross-modal fusion. Early fusion structure primarily performs simple concatenation operations [12] or element-wise addition [13] on the two types of features. Late fusion structure extracts the corresponding RGB and depth features, and then performs feature fusion at the end of the encoder. Zhou et al. [14] utilized a co-attention strategy between the encoder and decoder to fuse high-level features of both modalities. Zhou et al. [15] added the encoded RGB and depth features to construct a scale-aware module and selected appropriate scale features for each pixel in the decoding stage.

To improve feature extraction efficiency, some researchers focus on multi-level fusion to achieve interaction between feature extraction and feature fusion [46,47]. One approach treats the depth map as auxiliary information to enhance RGB features [20,21]. Chen et al. [22] constructed spatial deformation convolutions using depth features to enhance RGB feature encoding, requiring only a small number of additional parameters and computations. However, such methods did not fully extract features from depth maps. Differently, some researchers explore symmetrical structures to better capture the complementarity between RGB abd depth features. Chen et al. [19] introduced a separate and aggregate gate operation to jointly filter and recalibrate RGB and depth representations before the bottom-up cross-modal aggregation. Seichter et al. [48] performed multi-level attention fusion on depth features during the encoding stage, and enhanced the decoding capability using Non-Bottleneck-1D in the decoding stage. Wu et al. [49] proposed an interactive attention mechanism to fuse RGB and depth features to enhance the representation of the object of interest. However, these approaches did not fully consider the potential noise introduced by the low-quality depth map. Zhou et al. [50] proposed a bilateral cross-modal interaction network to capture cross-modal complementary cues. Zhao et al. [51] proposed a cross-modal attention fusion network to learn multi-modal and multi-level information by using coordinate attention feature interaction and gated cross-attention feature fusion.

Additionally, there are also alternative methods that deviate from the aforementioned structures. Lin et al. [52] utilized context-aware receptive fields and a zig-zag architecture to enhance the accuracy of feature aggregation and information propagation for two modalities. Wang et al. [53] took a different approach and chose to fuse multimodal information through channel exchanging between different modalities. Zhou et al. [54] proposed a novel uncertainty-aware transformer localization network to explore features from different angles.

In this paper, our EFDCNet utilizes a multi-level fusion approach in the network design. During the encoding stage, the encoding fusion module is employed to fuse the features extracted from RGB and depth maps, enabling cross-modality refinement and interaction. This effectively suppresses low-quality depth noise and enhances feature extraction capability. In the decoding stage, we leverage high-level information as guidance to refine the decoded features, resulting in the generation of high-resolution segmentation maps.

3. Proposed model

In this section, we elaborate on our EFDCNet. We first introduce the overview of our EFDCNet in Section 3.1. In Section 3.2 and Section 3.3, we give a detailed introduction of our Encoding Fusion Module (EFM) and Decoding Correction Module (DCM), respectively. At the end of this section, we clarify the loss function.

3.1. Network overview

As shown in Fig. 2, our EFDCNet takes RGB image and depth map as inputs, with the size of $3 \times 640 \times 480$ and $1 \times 640 \times 480$, respectively. It adopts the encoder-decoder structure, and consists of an RGB branch, a depth branch, the EFM, the DCM, and a context fusion module. In the encoder, both the RGB branch and the depth branch utilize ResNet [55] as the backbone for feature extraction. Each branch consists of five convolutional blocks denoted as D$_i$/R$_i$, and its output features are denoted as $F_D^i/F_R^i \in \mathbb{R}^{h_i \times w_i \times c_i}$ $(i = 1, 2, 3, 4, 5)$. Moreover, five EFMs in the encoder are used to incorporate depth features from each level into RGB features. The enhanced RGB features $\widehat{F}_R^i \in \mathbb{R}^{h_i \times w_i \times c_i}$ are passed into the next level. The decoder comprises a context fusion module [48] and three DCMs. $\widehat{F}_R^5$ from the last level of the encoder is being further enhanced by the context fusion module. The context fusion module is derived from ESANet [48] and is similar to the Pyramid Pooling Module in PSPNet [34]. In the context fusion module, we perform pooling operations with four different sizes of 1×1, 2×2, 4×4, and 8×8 to obtain multi-scale feature maps. Then, we reduce the channel number of these feature maps using 1×1 convolutions, upsample them using bilinear interpolation, and concatenate them along the channel dimension to generate the output features. The output features of the context fusion module serve not only the initial features F_{Dec}^4 of the decoder, but also the semantic information F_{Sem} for feature correction. In DCM, its

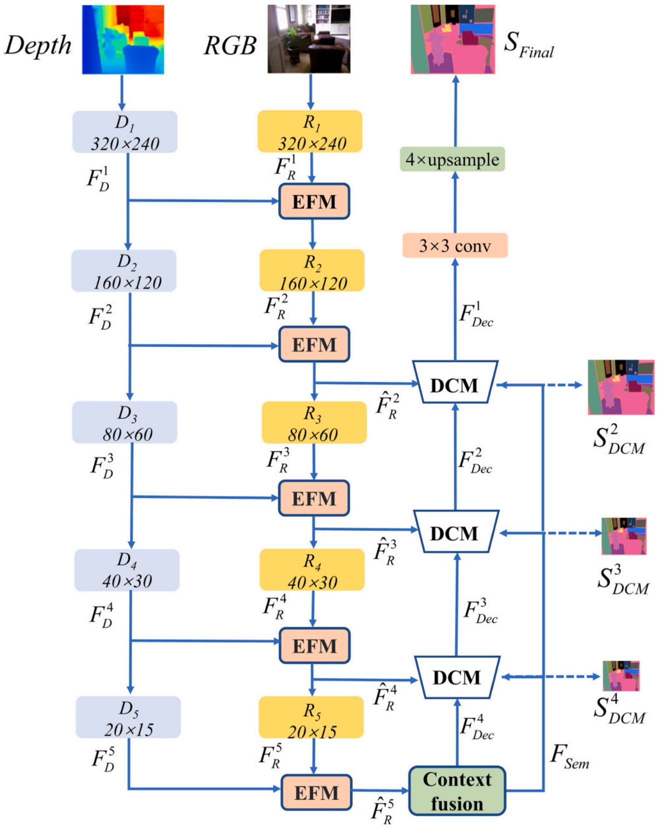

Fig. 2. The overall architecture of our EFDCNet.

inputs are F^i_{Dec} ($i = 4, 3, 2$), F_{Sem} and $\widehat{F}^i_R$, while its output is F^{i-1}_{Dec}. DCM uses F_{Sem} to filter out the noises in $\widehat{F}^i_R$ and correct the output features of the previous DCM. Through three DCM, our decoder produces the final semantic segmentation map S_{Final}. Here, we adopt a multi-scale supervision technique by adding semantic supervision at each DCM and S_{Final} for effective supervision.

3.2. Encoding fusion module

To address two issues in the encoder of existing methods, we design the EFM to improve the cross-modal feature fusion ability and alleviate the adverse effects of low-quality depth maps. Our EFM is in charge of incorporating the depth features into the RGB features, playing an important role in the encoder. Here, we provide a detailed description of the EFM from two aspects: feature purification and cross-modal interaction and fusion. We show the architecture of the EFM in Fig. 3.

1) Feature Purification. Attention mechanisms are highly effective in highlighting important information in computer vision. Inspired by SENet [30] and DCFNet [56], we propose a channel-wise filtering strategy that aims to minimize the impact of positional errors in the depth map and filter out redundant features. By employing this strategy, we can maximize the reduction of errors caused by incorrect depth map positions and effectively remove unnecessary features.

As shown in Fig. 3, the inputs of the EFM are $F^i{}_R$ and $F^i{}_D$. The channel attention maps reflect the importance of RGB and depth features, denoted as $Atta^i{}_R$ and $Atta^i{}_D$. We transfer the attention maps to the input features through the channel-wise multiplication to explicitly focus on important features and suppress unnecessary ones for scene understanding, generating the high-quality features $\dot{F}^i_{R/D}$. This process can be defined as:

$$Atta^i_{R/D} = \delta\left(W^i_{R/D}{}^{*}AvgPooling\left(F^i_{R/D}\right) + b^i_{R/D}\right), \tag{1}$$

$$\dot{F}^i_{R/D} = Atta^i_{R/D} \circledast F^i_{R/D}, \tag{2}$$

where W^i and b^i are the parameters of the fully connected layers, *AvgPooling*$(\cdot)$ represents global average pooling operation and δ represents the *Sigmoid* activation function, and $\circledast$ represents the channel-wise multiplication. Through purifying both modal features, the noise in the depth features is effectively suppressed, while the redundant features in the RGB features are filtered out.

In Fig. 4, we visualize the features to demonstrate the effectiveness of the feature purification. We can observe that F^2_D obtained through ResNet [55] is closer to the original depth map and retains a significant amount of noise. After the feature purification, the generated $\dot{F}^2_D$ becomes clearer, especially at the edges. This indicates that the weights of channel features affected by noise have decreased, leading to the suppression of overall feature map noise.

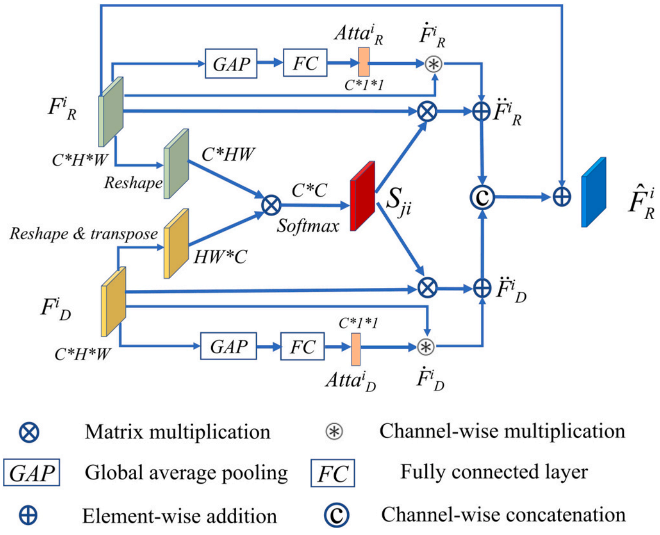

Fig. 3. Illustration of the encoding fusion module.

2) Cross-modal Interaction and Fusion. Previous strategies for cross-modal interaction mainly focus on local content aggregation between two modalities, where the information from the RGB image only establishes local dependencies with the depth map, without fully exploring the global dependencies between features. Inspired by the non-local model [31] and CANet [14], we utilize the self-attention mechanism to establish global dependencies between the depth features and RGB features, enabling effective extraction of discriminant and powerful features.

The input RGB and depth features undergo corresponding matrix transformation operations to generate new features $A \in \mathbb{R}^{C\times HW}$ and $B \in \mathbb{R}^{HW\times C}$, as illustrated in Fig. 3. They can be utilized to calculate a collaborative attention affinity matrix. Firstly, the collaborative attention affinity matrix $S \in \mathbb{R}^{C\times C}$ between RGB and depth features is obtained by matrix multiplication and Softmax layer:

$$S_{ji} = \frac{exp\left(A_i \times B_j\right)}{\sum_{j=1}^{C} exp\left(A_i \cdot B_j\right)}, i, j \in \{1, \ldots, C\}, \tag{3}$$

where S_{ji} represents the correlation between channel-level features from the i-th and j-th channels in different feature maps. Then we multiply the collaborative attention affinity matrix with the original RGB and depth features, generating $\ddot{F}^i_R$ and $\ddot{F}^i_D$ as follows:

$$\ddot{F}^i_{R/D} = S_{ji} \otimes F^i_{R/D}, \tag{4}$$

where $\otimes$ is the matrix multiplication. These enhanced features can effectively supplement the channel-wise filtering strategy in feature purification. We integrate $\dot{F}^i_{R/D}$ and $\ddot{F}^i_{R/D}$ through the element-wise addition. Finally, we fuse the above cross-modal features as follows:

$$\widehat{F}^i_R = Conv_{1\times1}\left(Concat\left(\left(\dot{F}^i_R + \ddot{F}^i_R\right), \left(\dot{F}^i_D + \ddot{F}^i_D\right)\right)\right) + F^i_R. \tag{5}$$

Through cross-modal interaction, the original RGB and depth features can utilize the global dependency relationship to effectively interacte with another modality. With feature purification and cross-modal interaction working together, the fused features can extract valuable information from cross-modal features, even from low-quality depth maps.

3.3. Decoding correction module

Low-resolution high-level semantic information are obained from the last-layer of encoder by the previous encoding feature extract and fusion. As mentioned earlier, the decoding stage involves the process of restoring the low-resolution semantic information to a high-resolution one. However, upsampling inevitably leads to the loss of global information, and the introduction of encoder features through skip connections may introduce noise. Our DCM addresses these problems by introducing a high-level semantic information branch to individually correct the upsampling information in the decoder and the low-level information in the encoder. In the following, we provide a detailed explanation of the DCM, which consists of the Compensation Unit and Filtering Unit.

1) Compensation Unit. As shown in Fig. 5, our DCM consists of three inputs, i.e., F^i_{Dec}, F_{Sem}, and $\widehat{F}^i_R$. In the Compensation Unit, the low-resolution F_{Sem} is adaptively integrated into each position of the upsampled decoder features F^i_{Dec}. This allows the unit to fuse global

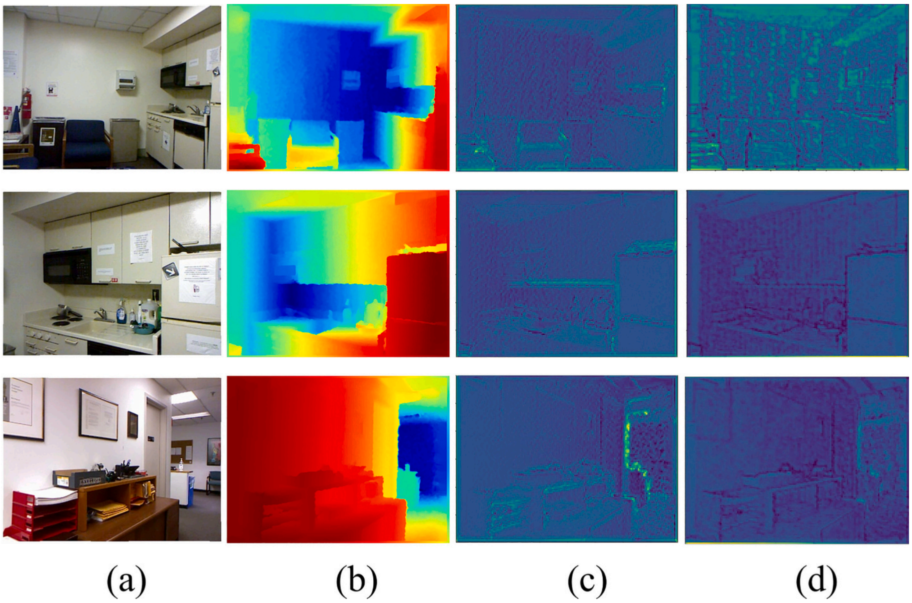

Fig. 4. The visualization of the features in the feature purification. (a) RGB image. (b) Depth map. (c) F_D^2. (d) $\dot{F}_D^2$.

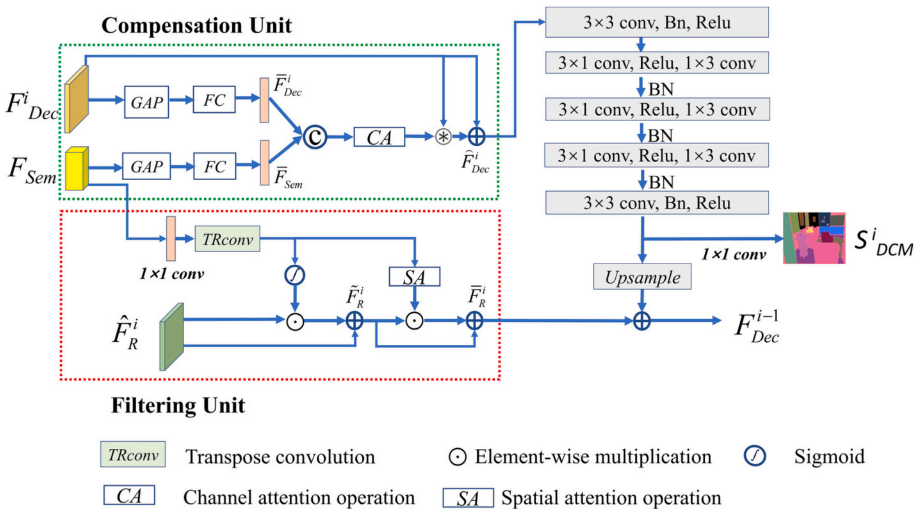

Fig. 5. Illustration of the decoding correction module.

semantic information at different scales and compensate for the potential loss of location information during the upsampling process.

Concretely, firstly, F^i_{Dec} and F_{Sem} undergo global average pooling and fully connected layer in the channel dimension, generating two sets of features with global statistical information, denoted as $\overline{F}^i_{Dec}$ and $\overline{F}_{Sem}$. These two features are then concatenated and further passed through a channel attention operation. The output is multiplied with the original F^i_{Dec}, and then connects to F^i_{Dec} through a residual connection, generating $\widehat{F}^i_{Dec}$. The above process can be represented as follows:

$$\widehat{F}^i_{Dec} = CA\left(Concat\left(\overline{F}^i_{Dec}, \overline{F}_{Sem}\right)\right)\circledast F^i_{Dec} + F^i_{Dec}, \tag{6}$$

where CA represents the channel attention operation, and ⊛ represents the channel-wise multiplication.

After compensation, the current feature $\widehat{F}^i_{Dec}$ undergoes subsequent operations of convolution, ReLU, and BN as shown in Fig. 5. The output $\widehat{F}^i_{Dec}$ is processed from two aspects, that is, one is processed to generate the predicted semantic segmentation map S^i_{DCM}, while the other is upsampled and added to the output of the filtering unit.

We visualize the features in DCM in Fig. 6 to validate the effectiveness of compensation units. By comparing Fig. 6(b) F^2_{Dec} and Fig. 6(c) $\widehat{F}^2_{Dec}$, we can observe that $\widehat{F}^i_{Dec}$ enhanced with the highest-level information F_{sem} exhibits increased activation values for foreground information, and each specific location of the objects of $\widehat{F}^2_{Dec}$ is highlighted. This demonstrates that the compensation units play a significant role in promoting the accuracy of object localization information in upsampled decoding process.

2) Filtering Unit. For the skip-connected encoded information, we design a Filtering Unit to utilize F_{Sem} to perform the spatial filtering twice on $\widehat{F}^i_R$. Firstly, we perform a 1 × 1 convolution and transpose convolution upsampling operation on the highest-level information to obtain $\widehat{F}_{Sem}$, ensuring that its spatial dimensions match the size of $\widehat{F}^i_{RGB}$. Then, we apply a sigmoid layer to $\widehat{F}_{Sem}$, leveraging its powerful integrated semantic information to construct a spatial gate map. This spatial gate map is then element-wise multiplied with the upsampled $\widehat{F}^i_R$ to filter

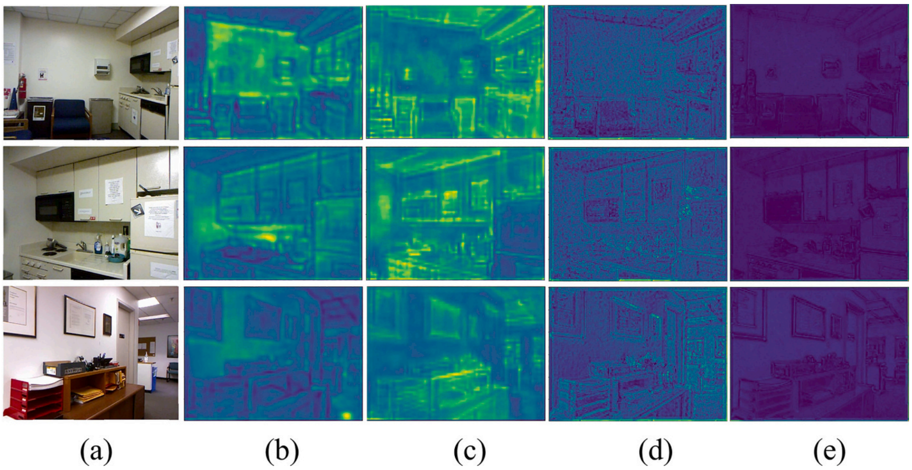

Fig. 6. The visualization of the features in DCM. (a) RGB image. (b) F^2_{Dec}. (c) $\widehat{F}^2_{Dec}$. (d) $\widehat{F}^2_R$. (e) $\overline{F}^2_R$.

out noises, generating $\widetilde{F}^i_R$. We name the above process as spatial gate filtering, which is formulated as follows:

$$\widetilde{F}^i_R = \widehat{F}^i_R \odot Sigmoid(\widehat{F}_{Sem}) + \widehat{F}^i_R, \tag{7}$$

where $\odot$ represents the element-wise multiplication.

After the spatial gate filtering, we construct a spatial attention operation to further filter out the noise from $\widetilde{F}^i_{RGB}$. By performing spatial attention operations, we can learn a robust spatial attention map, which highlights the regions of interest and suppresses interference from irrelevant regions. We filter $\widetilde{F}^i_R$ with the spatial attention map through the element-wise multiplication, and adopt a residual connection to preserve the original information. We represent the second spatial filtering as follows:

$$\overline{F}^i_R = \widetilde{F}^i_R \odot SA(\widehat{F}_{Sem}) + \widetilde{F}^i_R, \tag{8}$$

where SA represents the spatial attention operation. Finally, the filtered encoding feature $\overline{F}^i_R$ is added to the upsampled $\widehat{F}^i_{Dec}$, generating the output of DCM F^{i-1}_{Dec}.

By comparing Fig. 6(d) $\widehat{F}^2_R$ and Fig. 6(e) $\overline{F}^2_R$, it can be observed that after being filtered by highest-level information F_{sem}, the noise in $\widehat{F}^2_R$ obtained in the encoding phase has been significantly suppressed.

3.4. Loss function

As shown in Fig. 2, our EFDCNet adopts a multi-scale supervision strategy, that is, the three DCMs respectively output the predicted semantic segmentation maps S^4_{DCM}, S^3_{DCM}, and S^2_{DCM} with the size of 1/16, 1/8, and 1/4 of the original input size. The output of the last DCM, i.e., F^2_{Dec}, is used to generate the final semantic segmentation map S_{Final} through a convolution layer, and S_{Final} is upsampled by 4 times to restore its size to the same as the size of the ground truth. We calculate the cross-entropy loss between each of these four semantic segmentation maps and the corresponding ground truths, and then sum the above four losses for gradient backpropagation.

4. Experiments

4.1. Experimental setup

1) Datasets. We evaluate our method on two commonly used datasets:

NYU-Depth V2 dataset [63], also named *NYUDv2*, consists of 1449 RGB-D images with dense pixel annotations, including 795 training images and 654 test images. We employ labels that encompass 40 categories for evaluation.

SUN-RGBD dataset [64] comprises 10,335 RGB-D images with dense pixel annotations. Among them, 5285 images are allocated for training, while 5050 images are designated for testing. We utilize labels representing 37 categories for evaluation.

2) Implementation Details. Our method was constructed using the PyTorch framework [65], and all experiments were conducted on two NVIDIA 1080Ti GPUs. In the encoder, we chose the backbone from ResNet34, ResNet50, and ResNet101 [55]. We used SGD [66] as the optimizer with an initial learning rate ranging from 0.001 to 0.03. The epoch decay factor was set to 0.9, and the weight decay was set to 0.0005. During the training processing, we employed a cosine annealing strategy to adjust the learning rate. We trained for 500 epochs on the NYUDv2 dataset and 400 epochs on the SUN RGB-D dataset. For a comprehensive evaluation and comparison with other research, we additionally performed multi-scale testing. We used a scale range of (0.75, 1, 1.25, 1.75) and applied horizontal flipping to further enhance the robustness of our results.

3) Evaluation Metrics. We evaluate the performance using commonly used evaluation metrics in semantic segmentation, including Pixel Accuracy (PA), Class Mean Accuracy (CMA), and mean Intersection over Union (mIoU). PA measures the accuracy at the pixel level by dividing the number of correctly classified pixels by the total number of pixels in the image. CMA calculates the average accuracy for each class by dividing the number of correctly classified pixels for a specific class by the total number of pixels belonging to that class, providing insights into the model's performance on individual classes. mIoU measures the average overlap between the predicted segmentation masks and the ground truth masks for each class. It is calculated by dividing the intersection area of the predicted mask and ground truth mask by the union area of the two masks and provides an overall evaluation of the segmentation performance.

4.2. Performance analysis

To evaluate the effectiveness of our algorithm, we selected 17 advanced deep learning-based methods for comparison. We measured these methods from two aspects: quantitative comparison and visual comparison.

4.2.1. Quantitative comparison

We conduct quantitative comparisons with other methods on the NYUDv2 and SUN RGB-D datasets. In Table 1, we evaluate the performance of our EFDCNet using three different backbone networks:

Table 1
Quantitative comparison with state-of-the-art methods on the NYUDv2 test dataset. "*" represents the results of multi-scale testing.

Methods	Backbone	Input data	PA (%)	CAM (%)	mIoU (%)
ESANet [48] (2021, ICRA)	ResNet34	RGB-D	–	–	48.8
CMANet [17] (2022, Sensors)	ResNet50	RGB-D	73.9	59.8	47.3
RAFNet [57] (2021, Displays)	ResNet50	RGB-D	73.8	60.3	47.5
CMANet [17] (2022, Sensors)	ResNet50	RGB-HHA	74.2	60.2	47.6
RDFNet [58] (2017, ICCV)	ResNet50	RGB-HHA	74.8	60.4	47.7
SGNet [22] (2021, TIP)	ResNet50	RGB-D	75. 0	60. 8	47.7
SAMD [15] (2022, NC)	ResNet50	RGB-D	–	–	48.1
ACNet [23] (2019, ICIP)	ResNet50	RGB-D	74.9	61.2	48.3
Link-RGBD [49] (2022, SENS J)	ResNet50	RGB-D	76.8	59.6	49.5
CANet [14] (2022, PR)	ResNet50	RGB-D	75.9	63.9	50
ESANet [48] (2021, ICRA)	ResNet50	RGB-D	77.1	63.8	50.5
CMAFNet [51] (2023, NC)	ResNet50	RGB-D	76.1	64.2	50.5
NANet [18] (2021, SPL)	ResNet50	RGB-D	77.1	66.7	51.4*
RefineNet [59] (2017, CVPR)	ResNet101	RGB	72.8	57.8	44.9
LSD-GF [60] (2017, CVPR)	ResNet101	RGB-HHA	71.9	60.7	45.9
SCN [61] (2018, TCYB)	ResNet101	RGB-HHA	–	–	48.3
RGBXD [62] (2021, NC)	ResNet101	RGB-D	75	61.7	48.6
SGNet [22] (2021, TIP)	ResNet101	RGB-D	75. 6	61. 9	49.6
ShapeConv [12] (2021, ICCV)	ResNet101	RGB-D	75.8	62.8	50.2
CMAFNet [51] (2023, NC)	ResNet101	RGB-D	77.7	64.8	51.3
CANet [14] (2022, PR)	ResNet101	RGB-D	77.1	64.6	51.5*
NANet [18] (2021, SPL)	ResNet101	RGB-D	77.9	66.7	52.3*
SAMD [15] (2022, NC)	ResNet101	RGB-D	–	–	52.3*
SA-Gate [19] (2020, ECCV)	ResNet101	RGB-HHA	77. 9	–	52.4*
EFDCNet (Ours)	ResNet34	RGB-D	76.3	63.2	**49.8**
EFDCNet (Ours)	ResNet34	RGB-D	76.9*	64.2*	**50.8***
EFDCNet (Ours)	ResNet50	RGB-D	76.9	65.3	**51.4**
EFDCNet (Ours)	ResNet50	RGB-D	77.4*	65.4*	**51.9***
EFDCNet (Ours)	ResNet101	RGB-D	77.2	65.6	**52.0**
EFDCNet (Ours)	ResNet101	RGB-D	77.8*	65.7*	**52.7***

ResNet34, ResNet50, and ResNet101. For the single-scale test on ResNet50, our EFDCNet achieved PA: 76.9%, CMA: 65.3%, and mIoU: 51.4%. While the PA is slightly lower than ESANet by 0.2%, our EFDCNet outperforms ESANet [48] by 1.5% in CMA and 0.9% in mIoU. In multi-scale testing, although our PA and CMA are slightly lower than NANet [18] by 0.2% and 1.4% respectively, our mIoU is higher by 0.5%. Similar trends are observed for the ResNet101 backbone.

In Table 2, we performed the same evaluation on the SUN RGB-D dataset. The SUN-RGBD dataset is larger in scale and exhibits more diverse scenes. Our results on both single-scale and multi-scale testing with ResNet50 still demonstrate advantages compared to other methods.

Table 3 displays the per-class classification results after training on the NYUDv2 dataset using ResNet50. Our method outperforms others in 13 out of 40 classes, demonstrating the robustness of our approach, particularly for challenging and hard-to-classify categories. This highlights the capability of our method to handle classification tasks on highly imbalanced training data with different classes.

Table 2
Quantitative comparison with state-of-the-art methods on the SUN RGB-D dataset in 37 classes. "*" represents the results of multi-scale testing.

RGBD method	Backbone	Data	PA (%)	CAM (%)	mIoU (%)
CMANet [17] (2022, Sensors)	ResNet50	RGB-D	81.1	59.3	47.2
RAFNet [57] (2021, Displays)	ResNet50	RGB-D	81.3	59.4	47.2
ACNet [23] (2019, ICIP)	ResNet50	RGB-D	–	–	48.1
SGNet [22] (2021, TIP)	ResNet50	RGB-D	81.8	60.9	48.5
Link-RGBD [49] (2022, SENS J)	ResNet50	RGB-D	83.1	53.5	48.4
CANet [14] (2022, PR)	ResNet50	RGB-D	81.6	59.0	48.1
NANet [18] (2021, SPL)	ResNet50	RGB-D	82.0*	–	48.0*
CMAFNet [51] (2023, NC)	ResNet50	RGB-D	82.0	59.7	48.6
EFDCNet (Ours)	ResNet50	RGB-D	82.4	61.3	**48.8**
EFDCNet (Ours)	ResNet50	RGB-D	82.6*	61.5*	**49.2***

4.2.2. Computational complexity comparison

We conduct an analysis of the number of parameters, inference speed, and computational amount of our proposed EFDCNet with different backbone networks and other existing networks. Table 4 presents the results of all methods in the same environment. Compared to other methods with the same backbone, our EFDCNet has fewer parameters, slightly higher than the lightweight ESANet [48]. The inference speed of our EFDCNet is at the mid-range level. Our EFDCNet achieves performance similar to the lightweight ESANet with low computational amount. Based on the analysis of the number of parameters, inference speeds, and computational amount, we conclude that our EFDCNet achieves a good balance between efficiency and performance.

4.2.3. Visual comparison

To further demonstrate the superiority of our EFDCNet, we perform a visual comparison with other methods on the NYUDv2 dataset, as shown in Fig. 7. We show different scenes in the dataset, such as living rooms, kitchens, bedrooms, dining rooms, and offices. Compared with other methods, our segmentation maps are less affected by noise and show advantages in details and boundary segmentation. As can be seen from Fig. 7, the actual quality of the depth map is poor and interferes with the segmentation results. In this case, the phenomenon of misplaced confusion in our EFDCNet's results is greatly reduced, even if some small objects were not recognized.

4.3. Ablation study

We conduct thorough ablation experiments to assess the effectiveness of each module in our EFDCNet on the NYUDv2 dataset. Specifically, we conduct separate and joint contribution assessments of EFM and DCM. We also perform an evaluation and analysis of the constituent elements within the two modules. For all ablation experiments, we train variables using the same parameters and dataset settings as described in Section 4.1 and evaluate the relevant performance.

1. The effectiveness of EFM and DCM. We propose two modules, i.e., EFM and DCM, to achieve feature fusion and feature correction. To evaluate the individual contributions of these two modules, we used ResNet-34 as the backbone and provide four variants: 1) Baseline (i.e., in the encoding part, the depth map is directly added element-wise to the RGB features, while in the decoding part, basic upsampling and skip-connection addition are retained); 2) Baseline + EFM; 3) Baseline +

Table 3
Quantitative comparison of mIoU of each category with state-of-the-art methods on the NYUDv2 dataset. The top two results in each column are in red and blue.

Method	Wall	Floor	Cabinet	Bed	Chair	Sofa	Table	Door	Window	Bookshelf
ACNet [23]	80.0	87.4	58.6	71	64.1	61.3	45.6	43.2	47.2	43.9
SGNet [22]	80.4	85.7	60.4	71.8	59.5	62.7	46.5	42.3	49	45.8
CANet [14]	80.1	87.6	62.6	73.2	62.7	65.4	47.3	43	48.1	40.4
CMANet [17]	77.7	86.2	59.6	72.5	60.3	61.1	43.3	35.5	43.8	38.6
ESANet [48]	81.3	88.7	64.1	75.8	66.1	66.4	46.4	36.4	46.9	46.1
EFDCNet (Ours)	81.8	88.1	66.0	74.9	65.6	64.8	47.4	44.7	47.5	47.2

Method	Picture	Counter	Blind	Desk	Shelf	Curtain	Dresser	Pillow	Mirror	Mat
ACNet [23]	60.6	67.2	56.1	24.8	15.8	51.4	39.3	46.4	42.5	37.4
SGNet [22]	60.6	64.9	61.3	20.5	17.3	51.2	42.8	40.9	53.6	29.3
CANet [14]	60.1	69.7	60.6	22.7	12.6	61.9	48.4	48.3	47.1	36.3
CMANet [17]	60.9	62.5	56.1	21.7	10.0	56.1	50.1	46.4	45.8	37.2
ESANet [48]	63.1	69.8	63.8	26.4	19.6	64.2	56.7	53.4	59.2	40.2
EFDCNet (Ours)	64.6	71.4	62.7	26.7	20.0	66.5	50.9	50.0	56.5	32.9

Method	Cloths	Ceiling	Books	Refridg	TV	Paper	Towel	Shower	Box	Board
ACNet [23]	23.4	75.9	29.8	49.3	54.1	32.4	42.9	20.5	11.2	72.0
SGNet [22]	20.5	76.9	30.5	64.8	58.7	28.8	38.1	23.3	9.4	82.3
CANet [14]	22	78.7	35.0	55.1	54.1	32.2	44.5	48.9	12.7	77.8
CMANet [17]	21.1	75.3	33.1	55.1	63.3	30.1	40.1	32.1	14.3	62.5
ESANet [48]	26.8	77.6	30.5	56.5	51.9	32.8	45.2	46.1	12.1	41.9
EFDCNet (Ours)	28.5	80	32.9	58.1	52.8	33.1	45.0	45.8	11.3	73.0

Method	Person	Stand	Toilet	Sink	Lamp	Bathtub	Bag	Othstr	Othfurn	Otherprop
ACNet [23]	76.5	47.1	78.1	63.1	50.5	61.8	10.9	28.1	19.3	39.2
SGNet [22]	74.8	43.6	67.6	55.7	45.9	42.7	6.1	31.9	15.6	38.2
CANet [14]	79.6	38.6	74.3	65.2	49.9	53.6	10.6	30.2	20.5	38.9
CMANet [17]	77.3	40.8	70.9	58.9	47.9	57.3	13.6	31.2	19.1	38.1
ESANet [48]	70.8	49.7	73.9	65.6	54.6	55.3	4.1	34.2	22.7	40.1
EFDCNet (Ours)	79.3	49.8	83.1	65.2	54.1	60.4	4.9	30.7	18.9	40.8

Table 4
Results of computational complexity analysis.

RGBD method	Backbone	Parameter (M)	FPS	Speed (ms)	Flops (G)
PGDENet [47] (2022, TMM)	ResNet50	963.0	3.7	275.9	1126,4
CMANet [17] (2022, Sensors)	ResNet50	117.8	28.4	35.2	136.7
ACNet [22] (2019, ICIP)	ResNet50	116.6	27.1	36.9	126.7
CANet [14] (2022, PR)	ResNet50	105.4	35.7	28.1	126.6
ESANet [48] (2021, ICRA)	ResNet50	71.6	24.8	40.3	65.8
EFDCNet (Ours)	ResNet34	60.7	30.9	32.4	58.5
EFDCNet (Ours)	ResNet50	72.5	26.9	37.2	66.1
EFDCNet (Ours)	ResNet101	110.5	17.3	57.7	111.7

DCM; and 4) Baseline + EFM + DCM. We report the quantitative results in Table 5. “Baseline” only achieves a mIoU of 47.8%, which is 2.0% lower than the complete EFDCNet, indicating that our EFM can indeed improve segmentation accuracy. With the help of EFM or DCM, the second and third variants respectively improve performance relative to “Baseline”.

2. The effectiveness of each component of the EFM. The EFM aims to integrate encoding features that capture cross-modal information at various levels. We design and implement the entire encoding module for optimal performance. To verify its contribution, we compare EFM with three variants and two compared methods. The decoding stage uses the baseline design without both the compensation unit and the refinement unit.

The E+ variant replaces EFM with element-wise summation at all levels. To demonstrate the effectiveness of local design in EFM, we design two variants for comparison. The ED variant removes the cooperative matrix generation module and only uses the attention modules of RGB and depth channels to add up. The EF variant removes the self-attention modules of both modalities and retains only the cooperative attention module for correction and fusion. In addition, we include ESANet and SA-Gate as two fusion methods for comparison to further demonstrate the superiority of our fusion strategy.

Table 6 shows the evaluation results of three variants and two compared methods. Overall, the experimental results show that the performance of complete EFM exceeds those of fully replaced variants. Compared with a simple summation between the two modalities, EFM achieves a 2.3% PA, 2.2% CMA, and 2.7% mIoU improvement. Compared with the two variants, there is an improvement, which also proves the importance of local design. This indicates that only one part cannot achieve high performance of the combined strategy. Compared with the two compared fusion strategies, there is also an improvement, which confirms that the module improves the challenges mentioned earlier, including the weak cross-modal interaction capability and the impact of low-quality depth maps.

3. The effectiveness of each component of the DCM. The DCM aims to restore high-resolution semantic information to the maximum extent by using the decoder correction. For fairness, we still use the element-wise summation strategy in the encoding stage. Here, we provide three variants of the DCM: 1) D-variant directly removes all DCM operations and uses the original ESANet enhanced decoding directly in the decoding stage; 2) DC variant only uses the refinement module without processing the upsampling information; and 3) DE variant only uses the compensation module without processing the skip connection information in the encoding stage.

Table 7 shows the evaluation results of the DCM ablation study. We can see that with the help of DCM, PA, CMA, and mIoU are improved by 2.3%, 2.8%, and 2.6%, respectively, compared to the original module. Compared with the two variants DC and DE, there is also a 0.6% and 0.5% improvement in mIoU, which also demonstrates the superiority of our DCM.

4.4. *Failure cases*

Although our method has shown certain advantages compared to

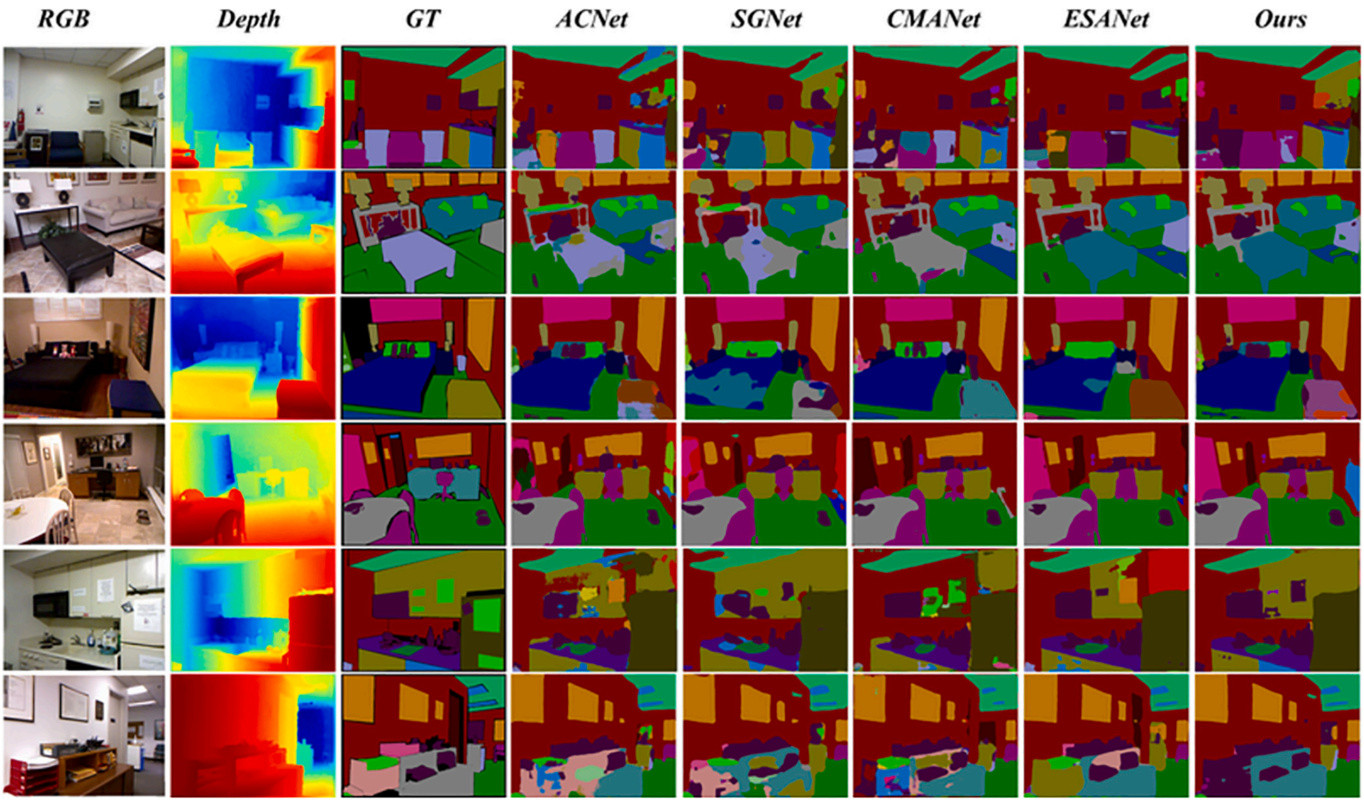

Fig. 7. Semantic segmentation visual comparision with state-of-the-art methods on the NYUDv2 dataset.

Table 5
Ablation experiments for EFM and DCM on the NYUDv2 dataset.

No.	Feature fusion	Feature correction	PA (%)	CAM (%)	mIoU (%)
1	average	No	73.5	59.9	47.8
2	EFM	No	74.9	62.9	48.5
3	average	DCM	75.8	61.7	49.1
4	EFM	DCM	76.3	63.9	49.8

Table 6
Ablation experiments for the effectiveness of each component of EFM and comparison with the other fusion methods on the NYUDv2 dataset.

No.	Variants	PA (%)	CAM (%)	mIoU (%)
1	E-	73.5	59.9	47.8
2	ED	74.8	61.5	48.6
3	EF	74.6	61.3	48.5
4	ESANet	75.1	60.8	48.2
5	SA-Gate	74.9	60.5	47.9
6	EFM (Ours)	75.8	62.1	49.2

Table 7
Ablation experiments for the effectiveness of each component of DCM on the NYUDv2 dataset.

No.	Variants	PA (%)	CAM (%)	mIoU (%)
1	D-	73.5	59.9	47.8
2	DC	74.9	61.2	48.5
3	DE	74.5	60.8	48.6
4	DCM (Ours)	75.8	61.7	49.1

similar approaches in the NYUDv2 and SUN RGB-D datasets tests, it performs poorly in three types of scenes sometimes. We show these failure cases in Fig. 8.

The first one is the unclear boundaries. Some adjacent objects at the same depth have challenging boundaries to capture. Inaccurate depth maps result in inaccurate pixel classification and segmentation results. For example, in the first case, there are two adjacent bins with unclear boundaries.

The second one is small objects and objects with weak textures. Small-sized objects occupy fewer pixels in the image, making segmentation more difficult. An example is the small lamp on the ceiling in the first case. Objects with weak semantic features are also challenging to segment, as they can be easily influenced by background noise. Examples include the doors in the first and second cases.

The third one is the complex scene. In some complex scenes, featuring overlapping, similar, or densely distributed objects. For example, there is too much overlapping of objects on the table in the third case, and there are too many interfering objects in the right area of the fourth case, resulting in poor segmentation results in both cases.

5. Conclusion

In this paper, we propose a novel and effective solution, named EFDCNet, for RGB-D indoor scene segmentation. For the encoding stage, EFM is proposed for extracting valuable information from depth maps with a channel-wise filter and enhancing cross-modal interactions with local information via self-attention, generating discriminant and powerful features. For the decoding stage, DCM is proposed for compensating for the upsampling information and filtering out the noises from the low-level encoding stage by using the highest-level information as semantic guidance, correcting the decoder features to boost accurate segmentation. With the collaboration of these two modules, our EFDCNet achieves high-performance semantic segmentation results on two datasets. Through performance analysis and ablation experiments, we demonstrate that our EFDCNet has certain advantages over other state-of-the-art methods in terms of quantitative comparison and visual comparison.

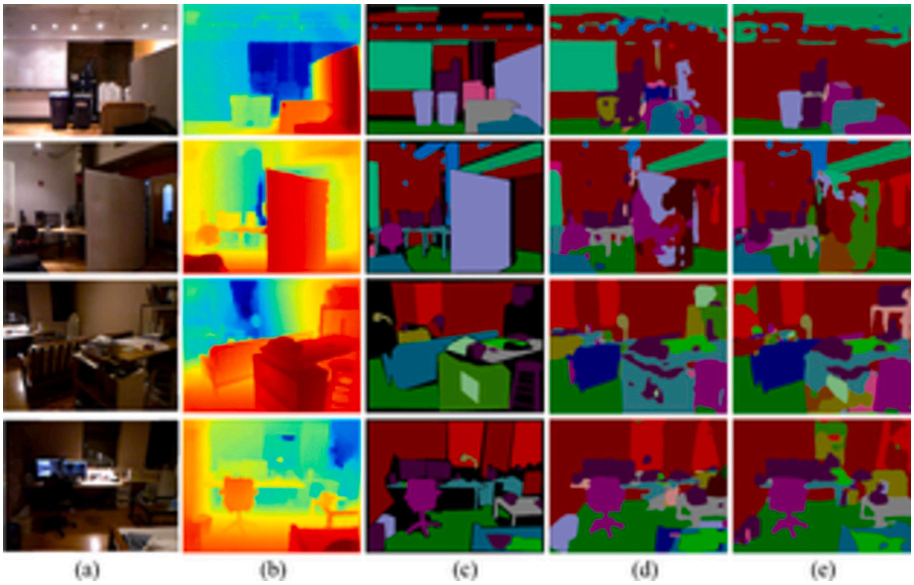

Fig. 8. Visual examples of failure cases. (a) RGB image. (b) Depth map. (c) GT. (d) Ours. (e) ESANet [48].

CRediT authorship contribution statement

Jianlin Chen: Conceptualization, Methodology, Software, Writing – original draft. **Gongyang Li:** Data curation, Writing – review & editing. **Zhijiang Zhang:** Visualization, Investigation, Formal analysis. **Dan Zeng:** Software, Validation.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

References

[1] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3431–3440.
[2] G. Li, Z. Liu, X. Zhang, W. Lin, Lightweight salient object detection in optical remote-sensing images via semantic matching and edge alignment, IEEE Trans. Geosci. Remote Sens. 61 (2023) 1–11.
[3] G. Li, Z. Liu, D. Zeng, W. Lin, H. Ling, Adjacent context coordination network for salient object detection in optical remote sensing images, IEEE Trans. Cybernet. 53 (1) (2023) 526–538.
[4] G. Li, Z. Bai, Z. Liu, X. Zhang, H. Ling, Salient object detection in optical remote sensing images driven by transformer, IEEE Trans. Image Process. 95 (2023) 5257–5269.
[5] D. Feng, C. Haase-Schütz, L. Rosenbaum, H. Hertlein, C. Glaeser, F. Timm, W. Wiesbeck, K. Dietmayer, Deep multi-modal object detection and semantic segmentation for autonomous driving: datasets, methods, and challenges, IEEE Trans. Intell. Transp. Syst. 22 (3) (2020) 1341–1360.
[6] B. Lewandowski, T. Wengefeld, S. Müller, M. Jenny, S. Glende, C. Schröter, A. Bley, H.-M. Gross, Socially compliant human-robot interaction for autonomous scanning tasks in supermarket environments, in: Proceedings of IEEE International Conference on Robot and Human Interactive Communication (RO-MAN), 2020, pp. 363–370.
[7] M. Zhu, Z. Sun, Z. Zhang, Q. Shi, T. He, H. Liu, T. Chen, C. Lee, Haptic-feedback smart glove as a creative human-machine interface (HMI) for virtual/augmented reality applications, Sci. Adv. 6 (19) (2020) eaaz8693.
[8] S. Izadi, D. Kim, O. Hilliges, D. Molyneaux, R. Newcombe, P. Kohli, J. Shotton, S. Hodges, D. Freeman, A. Davison, et al., Kinectfusion: real-time 3D reconstruction and interaction using a moving depth camera, in: Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology, 2011, pp. 559–568.
[9] W. Wei, M. Xu, J. Wang, X. Luo, Bidirectional attentional interaction networks for rgb-d salient object detection, Image Vis. Comput. 104792 (2023).
[10] C. Yao, L. Feng, Y. Kong, S. Li, H. Li, Double cross-modality progressively guided network for rgb-d salient object detection, Image Vis. Comput. 117 (2022) 104351.
[11] Y. Zhang, D. Sidibé, O. Morel, F. Mériaudeau, Deep multimodal fusion for semantic image segmentation: a survey, Image Vis. Comput. 105 (2021) 104042.
[12] J. Cao, H. Leng, D. Lischinski, D. Cohen-Or, C. Tu, Y. Li, Shapeconv: Shape-aware convolutional layer for indoor RGB-D semantic segmentation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 7088–7097.
[13] C. Hazirbas, L. Ma, C. Domokos, D. Cremers, FuseNet: incorporating depth into semantic segmentation via fusion-based CNN architecture, in: Proceedings of the Asian Conference on Computer Vision (ACCV), 2016, pp. 213–228.
[14] H. Zhou, L. Qi, H. Huang, X. Yang, Z. Wan, X. Wen, CANet: co-attention network for RGB-D semantic segmentation, Pattern Recogn. 124 (2022) 108468.
[15] F. Zhou, Y.-K. Lai, P.L. Rosin, F. Zhang, Y. Hu, Scale-aware network with modality-awareness for RGB-D indoor semantic segmentation, Neurocomputing 492 (2022) 464–473.
[16] A. Caglayan, N. Imamoglu, R. Nakamura, Mmsnet: Multi-modal scene recognition using multi-scale encoded features, Image Vis. Comput. 122 (2022) 104453.
[17] L. Zhu, Z. Kang, M. Zhou, X. Yang, Z. Wang, Z. Cao, C. Ye, CMANet: cross-modality attention network for indoor-scene semantic segmentation, Sensors 22 (21) (2022) 8520.
[18] G. Zhang, J.-H. Xue, P. Xie, S. Yang, G. Wang, Non-local aggregation for RGB-D semantic segmentation, IEEE Sign. Proc. Lett. 28 (2021) 658–662.
[19] X. Chen, K.-Y. Lin, J. Wang, W. Wu, C. Qian, H. Li, G. Zeng, Bi-directional cross-modality feature propagation with separation-and-aggregation gate for RGB-D semantic segmentation, in: Proceedings of the European Conference on Computer Vision (ECCV), 2020, pp. 561–577.
[20] Z. Wu, G. Allibert, C. Stolz, C. Ma, C. Demonceaux, Depth-adapted cnns for RGB-D semantic segmentation, arXiv (2022) preprint arXiv:2206.03939.
[21] Z. Wu, G. Allibert, C. Stolz, C. Demonceaux, Depth-adapted cnn for RGB-D cameras, in: Proceedings of the Asian Conference on Computer Vision (ACCV), 2020, pp. 388–404.
[22] L.-Z. Chen, Z. Lin, Z. Wang, Y.-L. Yang, M.-M. Cheng, Spatial information guided convolution for real-time RGB-D semantic segmentation, IEEE Trans. Image Process. 30 (2021) 2313–2324.
[23] X. Hu, K. Yang, L. Fei, K. Wang, ACNet: Attention based network to exploit complementary features for RGB-D semantic segmentation, in: Proceedings of the IEEE International Conference on Image Processing (ICIP), 2019, pp. 1440–1444.
[24] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: Proceedings of the International Conference on Medical Image Computing and Computer Assisted Intervention, 2015, pp. 234–241.
[25] V. Badrinarayanan, A. Kendall, R. Cipolla, SEGNet: a deep convolutional encoder-decoder architecture for image segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 39 (12) (2017) 2481–2495.
[26] F. Milletari, N. Navab, S.-A. Ahmadi, V-net: Fully convolutional neural networks for volumetric medical image segmentation, in: Proceedings of the Fourth International Conference on 3D Vision (3DV), 2016, pp. 565–571.
[27] Z. Tian, T. He, C. Shen, Y. Yan, Decoders matter for semantic segmentation: Data-dependent decoding enables flexible feature aggregation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3126–3135.
[28] J. Wang, K. Chen, R. Xu, Z. Liu, C.C. Loy, D. Lin, CARAFE: Content-aware reassembly of features, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3007–3016.

[29] Z. Zhang, X. Zhang, C. Peng, X. Xue, J. Sun, Exfuse: Enhancing feature fusion for semantic segmentation, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 269–284.

[30] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7132–7141.

[31] X. Wang, R. Girshick, A. Gupta, K. He, Non-local neural networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7794–7803.

[32] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A.L. Yuille, Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs, IEEE Trans. Pattern Anal. Mach. Intell. 40 (4) (2017) 834–848.

[33] T.-H. Tsai, Y.-W. Tseng, Bisenet v3: bilateral segmentation network with coordinate attention for real-time semantic segmentation, Neurocomputing 532 (2023) 33–42.

[34] H. Zhao, J. Shi, X. Qi, X. Wang, J. Jia, Pyramid scene parsing network, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2881–2890.

[35] H. Zhang, K. Dana, J. Shi, Z. Zhang, X. Wang, A. Tyagi, A. Agrawal, Context encoding for semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7151–7160.

[36] J. Fu, J. Liu, Y. Wang, Y. Li, Y. Bao, J. Tang, H. Lu, Adaptive context network for scene parsing, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 6748–6757.

[37] J. Li, S. Zha, C. Chen, M. Ding, T. Zhang, H. Yu, Attention guided global enhancement and local refinement network for semantic segmentation, IEEE Trans. Image Process. 31 (2022) 3211–3223.

[38] G. Li, Y. Wang, Z. Liu, X. Zhang, D. Zeng, RGB-T semantic segmentation with location, activation, and sharpening, IEEE Trans. Circuits Syst. Video Technol. 33 (3) (2023) 1223–1235.

[39] K. Wang, G. Zhang, H. Bao, Robust 3D reconstruction with an RGB-D camera, IEEE Trans. Image Process. 23 (11) (2014) 4893–4906.

[40] M.B. Shaikh, D. Chai, RGB-D data-based action recognition: a review, Sensors 21 (12) (2021) 4246.

[41] Y. Xiao, V.R. Kamat, C.C. Menassa, Human tracking from single rgb-d camera using online learning, Image Vis. Comput. 88 (2019) 67–75.

[42] G. Li, Z. Liu, L. Ye, Y. Wang, H. Ling, Cross-modal weighting network for RGB-D salient object detection, in: Proceedings of the European Conference on Computer Vision (ECCV), 2020, pp. 665–681.

[43] G. Li, Z. Liu, H. Ling, ICNet: information conversion network for RGB-D based salient object detection, IEEE Trans. Image Process. 29 (2020) 4873–4884.

[44] G. Li, Z. Liu, M. Chen, Z. Bai, W. Lin, H. Ling, Hierarchical alternate interaction network for RGB-D salient object detection, IEEE Trans. Image Process. 30 (2021) 3528–3542.

[45] X. Zhou, G. Li, C. Gong, Z. Liu, J. Zhang, Attention-guided rgbd saliency detection using appearance information, Image Vis. Comput. 95 (2020) 103888.

[46] J. Wang, Z. Wang, D. Tao, S. See, G. Wang, Learning common and specific features for RGB-D semantic segmentation with deconvolutional networks, in: Proceedings of the European Conference on Computer Vision (ECCV), 2016, pp. 664–679.

[47] W. Zhou, E. Yang, J. Lei, J. Wan, L. Yu, PGDENet: progressive guided fusion and depth enhancement network for RGB-D indoor scene parsing, IEEE Trans. Multimed. 25 (2022) 3483–3494.

[48] D. Seichter, M. Köhler, B. Lewandowski, T. Wengefeld, H.-M. Gross, Efficient RGB-D semantic segmentation for indoor scene analysis, in: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), 2021, pp. 13525–13531.

[49] P. Wu, R. Guo, X. Tong, S. Su, Z. Zuo, B. Sun, J. Wei, Link-RGBD: cross-guided feature fusion network for RGB-D semantic segmentation, IEEE Sensors J. 22 (24) (2022) 24161–24175.

[50] W. Zhou, Y. Yue, M. Fang, X. Qian, R. Yang, L. Yu, Bcinet: bilateral cross-modal interaction network for indoor scene understanding in rgb-d images, Inform. Fusion 94 (2023) 32–42.

[51] Q. Zhao, Y. Wan, J. Xu, L. Fang, Cross-modal attention fusion network for rgb-d semantic segmentation, Neurocomputing 548 (2023) 126389.

[52] D. Lin, H. Huang, Zig-Zag network for semantic segmentation of RGB-D images, IEEE Trans. Pattern Anal. Mach. Intell. 42 (10) (2019) 2642–2655.

[53] Y. Wang, F. Sun, W. Huang, F. He, D. Tao, Channel exchanging networks for multimodal and multitask dense image prediction, IEEE Trans. Pattern Anal. Mach. Intell. 45 (5) (2022) 5481–5496.

[54] W. Zhou, Y. Cai, L. Zhang, W. Yan, L. Yu, Utlnet: uncertainty-aware transformer localization network for rgb-depth mirror segmentation, IEEE Trans. Multimed. (2023) 1–11, https://doi.org/10.1109/TMM.2023.3323890.

[55] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.

[56] W. Ji, J. Li, S. Yu, M. Zhang, Y. Piao, S. Yao, Q. Bi, K. Ma, Y. Zheng, H. Lu, et al., Calibrated RGB-D salient object detection, in: Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 9471–9481.

[57] X. Yan, S. Hou, A. Karim, W. Jia, RAFNet: RGB-D attention feature fusion network for indoor semantic segmentation, Displays 70 (2021) 102082.

[58] S.-J. Park, K.-S. Hong, S. Lee, RDFNet: RGB-D multi-level residual feature fusion for indoor semantic segmentation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2017, pp. 4980–4989.

[59] G. Lin, A. Milan, C. Shen, I. Reid, RefineNet: Multi-path refinement networks for high-resolution semantic segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1925–1934.

[60] Y. Cheng, R. Cai, Z. Li, X. Zhao, K. Huang, Locality-sensitive deconvolution networks with gated fusion for RGB-D indoor semantic segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3029–3037.

[61] D. Lin, R. Zhang, Y. Ji, P. Li, H. Huang, SCN: switchable context network for semantic segmentation of RGB-D images, IEEE Trans. Cybernet. 50 (3) (2018) 1120–1131.

[62] J. Cao, H. Leng, D. Cohen-Or, D. Lischinski, Y. Chen, C. Tu, Y. Li, RGB× D: learning depth-weighted RGB patches for RGB-D indoor semantic segmentation, Neurocomputing 462 (2021) 568–580.

[63] N. Silberman, D. Hoiem, P. Kohli, R. Fergus, Indoor segmentation and support inference from RGB-D images, in: Proceedings of the European Conference on Computer Vision (ECCV) 7576, 2012, pp. 746–760.

[64] S. Song, S.P. Lichtenberg, J. Xiao, Sun RGB-D: A RGB-D scene understanding benchmark suite, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 567–576.

[65] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al., Pytorch: an imperative style, high-performance deep learning library, Adv. Neural Inf. Proces. Syst. 32 (2019).

[66] L. Bottou, Stochastic gradient descent tricks, in: Neural Networks: Tricks of the Trade: Second Edition, 2012, pp. 421–436.