IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 12 March 2023, accepted 6 April 2023, date of publication 11 April 2023, date of current version 14 April 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3266251

RESEARCH ARTICLE

A Novel Semantic Segmentation Algorithm for RGB-D Images Based on Non-Symmetry and Anti-Packing Pattern Representation Model

YUNPING ZHENG[1], YUAN XU[1], SHENGJIE QIU[1], WENQIANG LI[1], GUICHUANG ZHONG[1], AND MUDAR SAREM[2]

[1]School of Computer Science and Engineering, South China University of Technology, Guangzhou 510006, China
[2]General Organization of Remote Sensing, Damascus, Syria

Corresponding author: Yunping Zheng (zhengyp@scut.edu.cn)

This work was supported in part by the Natural Science Foundation of Guangdong Province of China under Grant 2017A030313349, Grant 2021A1515011517, and Grant 2023A1515011288; in part by the National Natural Science Foundation of China under Grant 61300134; and in part by the National Undergraduate Innovative and Entrepreneurial Training Program under Grant 202110561070 and Grant 202110561066.

ABSTRACT With the rapid development of deep learning technology, the accuracy of image semantic segmentation tasks has been greatly improved. However, indoor RGB-D semantic segmentation remains a challenging problem because of the complexity of indoor environments. The emergence of depth sensors makes depth information gradually used to improve the effect of semantic segmentation. The splicing of weights such as between the RGB features and the depth features which are used as the input features of the neural network can effectively improve the accuracy of the indoor semantic segmentation tasks. Most previous researches focused on improving the performance of semantic segmentation by adjusting the structure of convolutional neural network. These researches have either added attention mechanism or performed data augmentation on input features, but they didn't make full use of the boundary information and the texture information of the original RGB image. In this paper, we propose a semantic segmentation algorithm for RGB-D images based on Non-symmetry and Anti-packing pattern representation Model (NAM). The core idea of the proposed algorithm is to take the channel-wise concatenation of pre-segmentation labels provided by the traditional hierarchical image segmentation and RGB-D features as the input of the neural network so as to guide the semantic segmentation tasks. The extensive experiments are conducted on the popular indoor RGB-D semantic segmentation datasets. When compared with the state-of-art algorithms, the experimental results presented in this paper show that our proposed method has improved the performance of image semantic segmentation networks on several popular neural network architectures.

INDEX TERMS Deep learning, hierarchical image segmentation, image representation, NAM, RGB-D, and semantic segmentation.

I. INTRODUCTION

Semantic segmentation plays a significant role in computer vision research. It refers to identifying images at the pixel level, that is marking the object category to which each pixel in the image belongs. At present, semantic segmentation has been widely used in intelligent tasks such as auto-driving, robots perception, and medical image diagnosis [1], [2], [3], [4], [5]. In recent years, the image segmentation methods based on deep learning have developed rapidly. Since Shelhamer et al. [6] proposed the Fully Convolutional Neural networks (FCNs), the Convolutional Neural Networks (CNNs) have achieved impressive results in semantic segmentation tasks. Thus, they are widely applied in the field of semantic

The associate editor coordinating the review of this manuscript and approving it for publication was Andrea F. Abate.

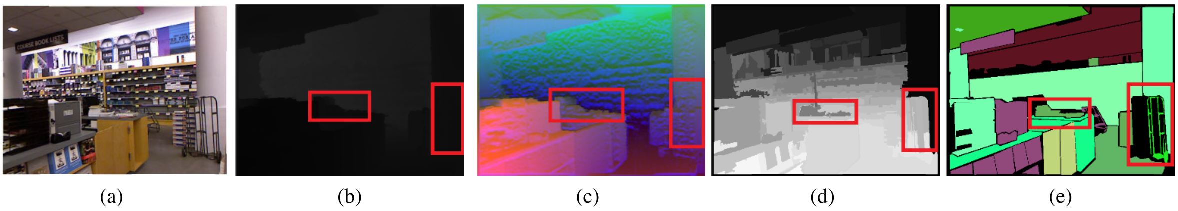

FIGURE 1. Semantic segmentation task of indoor scenes. (a) RGB raw image; (b) depth image; (c) HHA-encoded image; (d) NAMLab regions labels with 60th layers; (e) semantic segmentation labels.

segmentation. The researchers have proposed some famous neural network backbones such as AlexNet [7], VGGNet [8], and ResNet [9].

Semantic scene perception and understanding are two crucial tasks for mobile robots operating in various environments. However, the semantic segmentation of indoor scenes remains a challenging problem due to the complexity of indoor environments. For example, the changes in indoor lighting and occlusion between objects can easily cause a large number of pixels to be misclassified, thus affecting the final semantic segmentation result. With the widespread use of the depth sensors [10], the availability of the RGB-D data has driven the advances in RGB-D semantic segmentation. The RGB features describe the appearance of the information such as the object color and the brightness, while the depth features contain information related to the distance to the surface of the objects in the viewpoint scene. In recent years, some studies [11], [12], [13] combined the color information and the depth information as the input of the convolutional neural networks, hence they achieved good segmentation results.

In the indoor semantic segmentation researches, many studies in recent years have focused on mining the complementary information of the RGB color features and the depth features in the space and the shape, but these researches have ignored the inherent boundary features and the texture features of the indoor objects. Hence, they have failed to fully mine the complementary information between the texture, the shape and the color features in the original RGB image. However, a hierarchical image segmentation framework based on Non-symmetry and Anti-packing pattern representation Model (NAM) was proposed to divide an original image into layers with different regions according to various features of the image [14].

In the light of the above problem, by combining the RGB color features and the depth features, we propose a novel semantic segmentation algorithm for RGB-D images based on Non-symmetry and Anti-packing pattern representation model called NAMLab. The proposed algorithm firstly makes the image pixels merge into NAMLab blocks quickly and efficiently by defining the color difference of two pixels based on the Euclidean distance in Lab color space. Secondly, the algorithm defines the dissimilarity between two NAMLab-based regions and iteratively executing NAMLab-based merging algorithm of adjacent regions into larger ones to progressively generate a segmentation dendrogram. Finally, the hierarchical segmentation labels are extracted from the dendrogram. Moreover, in our proposed algorithm, the labels are added to the multi-channel input of the convolutional neural network. The hierarchical block labels generated by the NAMLab algorithm provide multi-layer boundary features of the objects in the image, which can effectively guide the semantic segmentation task. This strategy improves the performance of semantic segmentation tasks by adding channel-level features to the input of the neural network, which can be easily applied to most convolutional neural networks.

Fig.1 explains the semantic segmentation task of indoor scenes, from which the depth information ignores the inherent boundary features and the shape features of the objects themselves, while the proposed NAM region labels notice more details.

To verify the efficiency of our proposed technique, a thorough analysis has been conducted on the RGB-D image semantic segmentation benchmark NYUDv2 [15]. Meanwhile, we apply the proposed strategy to five popular semantic segmentation architectures. The main innovations of this paper are reflected in the following aspects:

Firstly, we propose a semantic segmentation strategy guided by NAM features to deal with indoor RGB-D semantic segmentation task.

Secondly, the NAM hierarchical features, the RGB features, and the depth features are channel-wise concatenated as the input of the neural network. By combining the advantages of the traditional image segmentation method with the advantages of the deep learning-based image semantic segmentation method, the performance of semantic segmentation tasks has been improved.

Finally, the experimental results presented show that the strategy proposed in this paper can effectively improve the performance of semantic segmentation tasks and also show better segmentation results on the popular RGB-D image datasets. In addition, the proposed strategy can be easily applied to different semantic segmentation networks.

The rest of the paper is organized as follows. Section II presents the related works about the image semantic segmentation and hierarchical image segmentation. In section III, a novel semantic segmentation algorithm for RGB-D images based on NAMLab model is described in details. Section IV presents the experimental results of our algorithm compared with the state-of-the-art algorithms. The conclusion and the possible future work are presented in section V.

II. RELATED WORKS

A. IMAGE SEMANTIC SEGMENTATION BASED ON CONVOLUTIONAL NEURAL NETWORK

Since Shelhamer et al. [6] proposed the FCN, the convolutional neural networks have been widely used in the tasks of semantic segmentation. The FCN replaces the fully connected layer in the classification network with a convolutional layer, hence the FCN can input an original image with any size, and then generate an output of the corresponding size through up-sampling. However, the convolution and pooling operations of the CNN greatly reduce the resolution and the size of the original image, resulting in loss of details. In response to this problem, Liu et al. [16] proposed the ParseNet, which improved the FCN by adding the context modules to obtain global information. Also, based on the FCN, Ronneberger et al. [17] proposed the U-Net. The U-shaped network structure down-samples four times in the encoder, and up-samples four times accordingly in the decoder. In the same stage, it adopts a skip connection structure instead of direct supervision on high-level semantic features. The high-level semantic feature map is restored to the resolution of the original image, thereby simultaneously acquiring the contextual information and the location information. Zhao et al. [18] proposed the Pyramid Scene Parsing Network (PSPNet), which uses a feature extractor with a dilated network strategy to extract the patterns from the input images and aggregate the contextual information from different regions. In the field of image semantic segmentation, one of the most famous frameworks is the DeepLab series. The DeepLab [19] used the atrous convolution kernels to avoid the information loss caused by max-pooling and down-sampling in the Deep Convolutional Neural Network (DCNN). It adopted the Conditional Random Field (CRF) to improve the model's ability of capturing the fine details. Then, the DeepLab V2 [20] used the Atrous Spatial Pyramid Pooling (ASPP), which in turn used multiple atrous convolutions with different sampling rates to integrate multi-scale features. Also, the ResNet was replaced with VGG16 in the DeepLab V2 and multiple astrous convolution kernels of different sampling rates were used to extract features. The DeepLab V3 [21] removed the CRF, improved the ASPP module, and reviewed atrous convolution. It expanded the receptive field to extract multi-scale information under the cascade module and the pyramid pooling framework. The DeepLab V3+ [22] further extended DeepLab V3 by adopting the Xception model in the semantic segmentation task and using the Spatial Pyramid Pooling (SPP) module in the encoder-decoder structure. The encoder gradually reduced the feature map to extract rich semantic information, and the decoder restored it.

B. HIERARCHICAL IMAGE SEGMENTATION

Compared with the image segmentation methods based on deep learning, the traditional image segmentation methods payed more attention to mining the color difference between sub-regions, the complementary information, and the hierarchical information between the texture and the shape features. In fact, the traditional image segmentation methods can effectively make up for the shortcomings of deep learning methods. There is now broad agreement that the performance of deep learning based segmentation algorithms is plateauing, especially in certain application domains such as medical image analysis [23]. To advance to the next level of performance, the authors in [23] thought that there is a need to further explore the combination of the CNN-based image segmentation models with the prominent "classical" model-based image segmentation methods. The integration of the CNNs with the graphical models has been studied, but the integration with active contours, graph cuts, and other segmentation models was fairly recent and deserved further work [23].

The popular Mean-Shift algorithm (MShift) [24], the Graph-based Image Segmentation algorithm (GBIS) [25] and the Multiscale Normalized Cut algorithm (MNCut) [26] were actually looking for an optimal segmentation of a given image. Although many new image segmentation algorithms have been proposed, how to effectively segment an image into regions that are "meaningful" to the human visual perception and ensure that the segmented regions are consistent at different resolutions is still a very challenging task up to now.

Most of the previous traditional image segmentation methods can only generate a single segmentation result. However, some researchers believe that multi-layer segmentation results with different target segmentation numbers may be able to better segment images. Syu et al. [27] put forward a hierarchical segmentation framework based on iterative contraction and merging. They stated that an algorithm that can only generate unique segmentation result may not be a suitable approach. Arbeláez et al. [28] put forward a gPb-OWT-UCM algorithm for hierarchical image segmentation. The algorithm first calculated the possibility of each pixel as a boundary through gPb. Then, the OWT transformed the gPb result into multiple closed regions. Finally, the UCM converted the above regions set into a hierarchical tree.

Inspired by the idea of the NAMLab and the "global-first" invariant perceptual theory, Zheng et al. [14] proposed a totally different framework for hierarchical image segmentation as they did not need to use the definition of the affinity value which was usually used in the energy functions and the graph Laplacian matrices. Also, they put forward a fast NAMLab-based algorithm for hierarchical image segmentation which, however, was also a traditional segmentation method.

C. IMAGE SEMANTIC SEGMENTATION BASED ON FUSION OF DEPTH FEATURES AND RGB FEATURES

With the wide application of depth sensors, researchers can easily obtain the depth information in the scene. The research on RGB-D images has also made great progress. In the current RGB-D image semantic segmentation tasks, the researchers have devoted themselves to three types of methods. The first type is to propose a strategy to fuse the depth features and the RGB features. The second type is to design a dedicated network architecture for the RGB-D data [13], [29]. And the third type is to design the structure of augmenting or replacing convolutional layers [30], [31]. However, our proposed strategy in this paper belongs to the first category.

As for the first category, Couprie et al. [32] proposed a pre-fusion method, which was a channel-level stitching of the RGB features and the depth features of the image as the input of the convolutional neural network. Gupta et al. [33] proposed an encoding method which converted a single-channel depth image to three-channel images by extracting Horizontal disparity, Height above ground, and the Angle of the pixel's local surface normal (HHA). The fusion could better guide indoor semantic segmentation task. The FuseNet [34] and the RedNet [35] fused the deep features into the RGB encoders, which followed the intuition that using the complementary depth information could further enhance the semantically richer RGB features.

As for the second category, Jiang et al. [35] proposed a gate fusion method that fused multi-level features from the backbone stage. Fooladgar and Kasaei [29] proposed an efficient encoder-decoder model with an attention-based fusion block to integrate the interaction between the feature maps of the depth mode and the RGB mode. Hu et al. [13] proposed an attention complementary network to selectively collect the features from the RGB and the depth branches.

As for the third category, Chen et al. [36] proposed the depth-aware convolutions based on the handcrafted Gaussian functions to weight pixels by exploiting the depth similarity between them. Cao et al. [37] designed a shape-aware convolutional layer that can replace the ordinary convolutional layer in semantic segmentation, making the network pay more attention to shape information when necessary and improving the performance of the RGB-D semantic segmentation tasks.

III. NAM-BASED SEMANTIC SEGMENTATION ALGORITHM FOR RGB-D IMAGES

In this section, we briefly describe the NAM method. Then, the hierarchical image features based on the NAM are presented. Finally, a NAM-based semantic segmentation algorithm for RGB-D images is put forward.

A. DESCRIPTION OF THE NAM

The Non-symmetry and Anti-packing pattern representation Model (NAM) [38], [39] is an anti-packing problem. The idea of the NAM can be briefly described as follows: Giving a packed pattern and n predefined sub-patterns with different shapes, pick up these n sub-patterns from the packed pattern, then represent the packed pattern with the combination of these sub-patterns.

The idea of the NAM can be described as follows: Giving a packed pattern and predefined sub-patterns with different shapes, pick up these sub-patterns from the packed pattern, then represent the packed pattern with the combination of these sub-patterns. The following are an abstract description of the NAM. Suppose the original pattern is Γ, the reconstruction pattern is Γ'. Then, the NAM is a transform model from Γ to Γ'. The procedure of the transform can be written as follows:

$$\Gamma' = T(\Gamma) \tag{1}$$

where $T(\cdot)$ is a transform or encoding function. The procedure of encoding can be obtained by the following expression:

$$\Gamma' = \cup_{j=1}^{n} p_j(v, A|A = \{a_1, a_2, \ldots, a_{m_i}\}) + \epsilon(d) \tag{2}$$

where Γ' is the reconstruction pattern; $P = p_1, p_2, \ldots, p_n$ is a set of some predefined sub-patterns; n is the number of sub-pattern types; $p_j \in P$ is the j^{th} sub-pattern $(1 \leq j \leq n)$; v is the value of p_j; and $A = \{a_1, a_2, \ldots, a_{m_i}\}$ is a parameter set of the sub-pattern $p_j(1 \leq j \leq n)$. If the types of two sub-patterns are different, the numbers and the meanings of parameters in A are different.

B. HIERARCHICAL IMAGE FEATURES BASED ON THE NAM

According to the human visual characteristics, in order to represent color images in perceptual uniformity, the NAMLab-based feature representation incorporated more robust local and global characteristics of images [14]. The proposed feature representation approach embraced the color, the spatial, the size, and the texture features to improve its processing ability in dealing with different image instances. The merging rule of NAMLab-based regions includes three modules, namely, the representation module, the merging module, and the removal module.

In the presentation module, the idea of the model is to represent the blocks of the input image by the mode of asymmetric inverse layout. An image is scanned line by line through raster scanning, and the distances between the adjacent pixels are judged according to the Lab color and the Gouraud formulas in order to expand the region, so that the original image is divided into one initial NAMLab rectangular region. Finally, the block map two-dimensional vector is used to record the NAMLab rectangular region number corresponding to each pixel, and its Lab feature mean and variance are also recorded.

In the merging module, for two adjacent NAMLab regions, when the difference between the mean values and the variance of the two Lab features and are less than two certain thresholds respectively, these two NAMLab blocks can be merged. The general process is as follows: Scan each NAMLab block in a raster way. For the current NAMLab block, first scan the NAMLab blocks of all adjacent pixels from the bottom to the top starting from the left side of the western border. If the NAMLab block to which the current adjacent pixel belongs is different from the current NAMLab block, then it is judged whether to merge the two NAMLab blocks judging by whether the dissimilarities between the NAMLab-based region is more than a certain thresholds, which will be described later. Then, scan all adjacent pixels from left to right starting from the northern boundary. If the NAMLab block to which the current adjacent pixel belongs is different from the current NAMLab block, it is judged whether to merge the two NAMLab blocks according to the dissimilarities between the NAMLab-based region. This scanning is repeated until all adjacent pixels starting from the northern boundary have been processed.

During the merging process of the NAMLab blocks, there are some small residual regions whose color mean and variance are quite different from the color mean and variance of their adjacent rejoins so that they cannot be merged. Therefore, we customize a threshold for the size of the region. When the size of the current region is smaller than the threshold, the current region will be merged into a region with the least difference among all the neighboring regions.

By taking the region obtained by the above method as the bottom node, the adjacent regions can be merged into a larger region, thereby gradually forming a dendrogram for hierarchical segmentation. Finally, the hierarchical segmentation results can be obtained. Fig.3 shows the entire flow of hierarchical image segmentation based on the NAMLab. The different colors in each segmentation map represent the different regions. The original image passes through the representation module, the merge module, the removal module, and the scanning module. Finally, the layered image features with different region numbers, such as 10, 20, 30, 40, 50 and 60 are output on the last column of Fig.3 from the top to the bottom, respectively.

To measure the dissimilarities between the NAMLab-based region i and region j, some formulas should be defined first and as they are described below. The dissimilarity measure of the region size between two regions is defined as follows:

$$S_{ij} = \frac{n_i n_j}{n_i + n_j} \tag{3}$$

where n_i and n_j denote the total number of the pixels in region i and region j, respectively.

The dissimilarity measure of the texture feature between two regions is defined as follows:

$$T_{ij} = ||w_i - w_j||_2 \tag{4}$$

where wld_i and wld_j represent the texture feature vectors of region i and region j, which are obtained according to the theory of Weber Local Descriptor [40].

The dissimilarity measure of the color feature between two regions is defined as follows:

$$M_{ij} = ||c_i - c_j||_2 \tag{5}$$

where c_i and c_j are the average LAB color of region i and region j, respectively.

To define the dissimilarity measure between the averaged color difference across the border of two regions, a 3×3 local window over the border region in the image is applied.

$$B_{ij} = \frac{\sum_{p \in br_i} \sum_{q \in br_j} ||c_p - c_q||_2}{N_{pq}} \tag{6}$$

where br_i denotes the areas which are the intersections of region i and the border area, so $p \in br_i$ and $q \in br_j$ denote regions on the two sides of the border.

For a pixel p_i, which represents a pixel in region i, check an 5×5 local window w_{p_i} at p_i and find the most common index in it, denoted as I_{p_i}. The dissimilarity measure of the spatial intertwining between two regions is defined as follows:

$$I_{ij} = min\left(\sum_{p_i} \psi\left(I_{p_i}, j\right), \sum_{q_j} \psi\left(I_{q_j}, i\right)\right) \tag{7}$$

where the function $\psi(\cdot)$ is defined as:

$$\psi(x, y) = \begin{cases} 1, & if x = y, \\ 0, & otherwise. \end{cases} \tag{8}$$

The formula D_{ij} describes the dissimilarity between the NAMLab-based region i and region j, which defines a more informative and comprehensive dissimilarity between the two regions i and j as follows:

$$D_{ij} = S_{ij} \times \frac{(\alpha M_{ij} + \beta T_{ij} + \gamma B_{ij})}{\sqrt{\lambda + I_{ij}}} \tag{9}$$

Regarding the selection of parameters α, β, and γ, they are the weights of color features, texture features, and edge features, respectively, in the task of measuring the differences between two NAMLab regions, While λ is the correction factor for the measure of the spatial intertwining. Their selection is based on experimental experience, where the optimal values are found by continuously adjusting the values of the four parameters according to the effect of image segmentation. Specifically, α=1.0, β=1.97, γ=1.97, λ=67.0.

C. NAM-BASED SEMANTIC SEGMENTATION STRATEGY FOR RGBD IMAGES

With the popularity of the depth sensor, the researchers are increasingly using the depth features of the images to guide the semantic segmentation tasks. On the basis of depth images, Gupta et al. [33] proposed an HHA encoding method that used three channels of each pixel to encode the depth image which are: The horizontal disparity, the height above ground, and the Angle of the pixel's local surface normal.

Fig.2 shows the RGB raw image, the depth image, the HHA feature image, and the NAMLab hierarchical feature image.

Different from the network structure specially designed for the RGB-D semantic segmentation, the strategy based on the NAMLab hierarchical feature guidance proposed in this paper is a more general method, thus it can be easily applied to the input of most convolutional neural networks, and be not limited to the RGB-D semantic segmentation tasks.

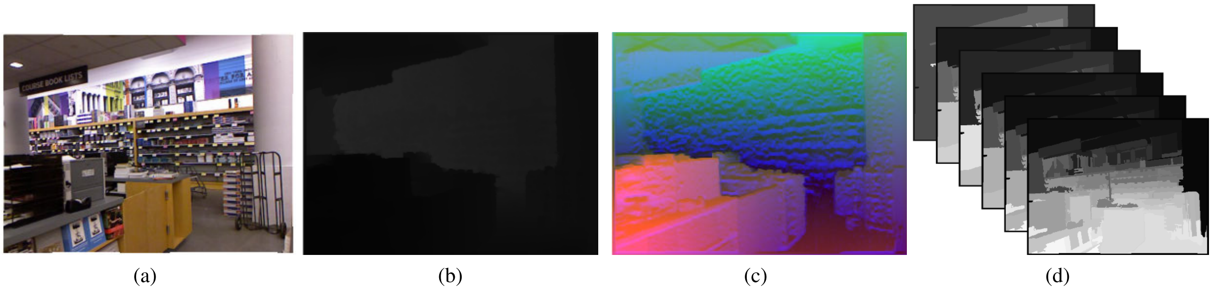

FIGURE 2. Original image and its three feature images. (a) an original RGB image; (b) its depth image; (c) its HHA depth image; (d) its NAMLab hierarchical segmentation image.

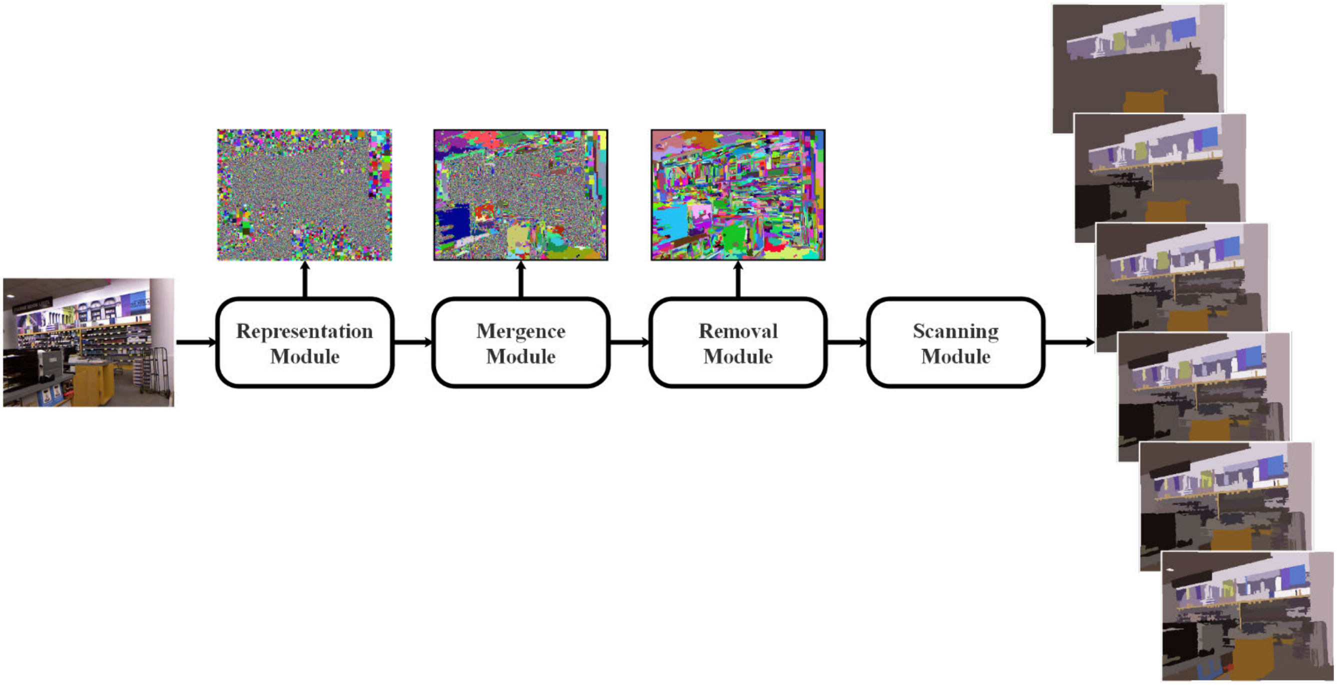

FIGURE 3. The entire flow of hierarchical image segmentation based on the NAMLab.

Fig.4 describes the entire strategy. In order to take advantage of the high-level backbones in semantic segmentation, we need to concatenate on the RGB raw image, the depth image, and the NAMLab hierarchical feature image in the channel dimension to be the input of the convolutional neural network. The depth image may be a Depth image or an HHA image and the network architecture takes FCN as an example. The symbol C indicates channel-level concatenation and the symbol OR indicates that the input depth information is either the HHA feature or the Depth feature.

IV. EXPERIMENTS

In this paper, the proposed strategy is implemented on the open-source deep learning framework called Pytorch. Except for the Baseline experiments refer to the experimental data of other researchers, all our experiments are carried out in the same hardware and software environment. The GPU uses a NVIDIA TITAN Xp. The CPU is Inter(R) Xeon®CPU E5-2680 v4 @ 2.40GHz. The capacity of the memory is 16GB. To verify the effectiveness of the proposed method, we evaluate our method on a popular RGB-D indoor image dataset NYUDv2 [15] and conduct ablation experiments. The NYUDv2 dataset contains 1449 indoor RGB-D images where 795 images are used for training and 654 images are used for testing. All the pixels of the images, for both training and testing, in this dataset are labeled as 13 classes (i.e., NYUDv2-13) and 40 classes (i.e., NYUDv2-40), respectively.

Our experimental results are evaluated with the following protocol and metrics. For the sake of explanation, we remark the following notation details [41]: Assuming a total of $k+1$ classes (from L_0 to L_k including a void class or background) and p_{ij} is the amount of pixels of class i inferred to belong to class j.

The Pixel Acc(PA) represents the predicted correct pixel value as a percentage of the total pixel value and it is defined

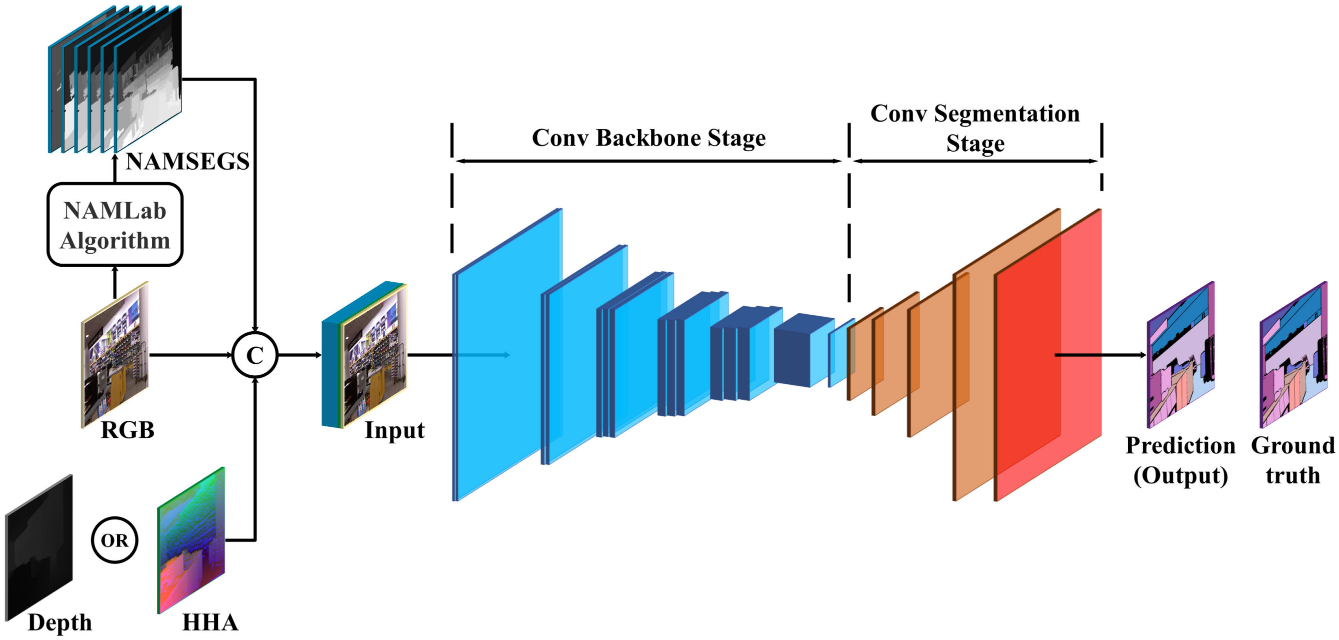

FIGURE 4. **Overall structure of RGB-D image semantic segmentation strategy based on the NAM feature guidance. The network architect takes FCN as example.**

as follows:

$$PA = \frac{\sum_{i=0}^{k} p_{ii}}{\sum_{i=0}^{k}\sum_{j=0}^{k} p_{ij}} \tag{10}$$

The Mean Acc(MPA) represents the average of the sum of the pixel accuracies of all classes and it is defined as follows:

$$MPA = \frac{1}{k+1}\sum_{i=0}^{k} \frac{p_{ii}}{\sum_{j=0}^{k} p_{ij}} \tag{11}$$

The Mean IoU($MIoU$) represents the average of the ratios between the intersection and the union of all classes predictions and the ground truth and it is defined as follows:

$$MIoU = \frac{1}{k+1}\sum_{i=0}^{k} \frac{p_{ii}}{\sum_{j=0}^{k} p_{ij} + \sum_{j=0}^{k} p_{ji} - p_{ii}} \tag{12}$$

Finally, the Fw IoU($FWIoU$) is an improved mean over the raw Mean IoU, which weights each class importance depending on its appearance frequency, and it is defined as follows:

$$FWIoU = \frac{1}{\sum_{i=0}^{k}\sum_{j=0}^{k} p_{ij}}\sum_{i=0}^{k} \frac{\sum_{j=0}^{k} p_{ij}p_{ii}}{\sum_{j=0}^{k} p_{ij} + \sum_{j=0}^{k} p_{ji} - p_{ii}} \tag{13}$$

We initialize the backbone using ResNet [9] and ResNeXt [42] models pre-trained on ImageNet [39]. The DeepLabV3+ is used as the main semantic segmentation network architecture as the baseline method. For all the baseline methods, the inputs are both channel-wise concatenation of the RGB raw images and the HHA depth images. Compared with the baseline methods, we only change the feature type of the inputs (i.e. the channel-wise concatenation of the NAMLab hierarchical features with the original image and the depth image) without making any change to other settings, thus guaranteeing that the performance improvement obtained is due to our proposed method, and not due to other factors. Besides, unless otherwise noted, strategies such as MultiScale-Flip (MS-F), Conditional Random Field CRF [19] or data augmentation are not employed in all our experiments.

On the NYUDv2-40 dataset, we have also conducted experiments on multiple popular semantic segmentation network architectures, such as DeepLabV3+, DeepLabV3, UNet, PSPNet and FPN. The backbone adopts the ResNet101 model pre-trained on ImageNet. On the same dataset, we have also conducted ablation experiments where the network architecture is DeepLabV3+, and the backbone remains the ResNet101 model pre-trained on ImageNet.

A. NAM-BASED SEMANTIC SEGMENTATION STRATEGY FOR RGBD IMAGES

The results of the baseline methods and our method using different backbones on NYUDv2-13 are shown in Tab.1. The architecture adopted is DeepLabV3+. In the Tab. 1, NAM-6 means that we have performed channel-wise concatenation of six layers of the NAMLab hierarchical features, the RGB original images and the HHA depth images, then we have

TABLE 1. Performance comparison with baselines on NYUDv2-13 dataset.

Back bone	Settings	Pixel Acc	Mean Acc	Mean IoU	Fw IoU
ResNet 50	Baseline	80	72.5	60.8	67.6
	NAM-6	80.2	73	61.2	67.8
	+	0.2	0.5	0.4	0.2
ResNet 101	Baseline	80	73.4	61.3	67.6
	NAM-6	81	74.2	62.7	69
	+	1	0.8	1.4	1.4
ResNext 101	Baseline	81.8	73.9	63.2	70.1
	NAM-6	81.7	75.3	63.6	69.9
	+	-0.1	1.4	0.4	-0.2

TABLE 2. Performance comparison with baselines on NYUDv2-40 dataset.

Backbone	Settings	Pixel Acc	Mean Acc	Mean IoU	Fw IoU
ResNet 50	Baseline	73.1	57.7	45.6	59.2
	NAM-1	73.3	57.8	45.3	59.4
	+	0.2	0.1	-0.3	0.2
	NAM-6	73.5	58.4	45.7	59.6
	+	0.4	0.7	0.1	0.4
ResNet 101	Baseline	73.4	58.9	45.9	59.7
	NAM-1	75	60.7	47.9	61.5
	+	1.6	1.8	2	1.8
	NAM-6	74.4	59.7	47.3	60.8
	+	1	0.8	1.4	1.1
ResNext 101	Baseline	74.7	61.5	48.9	61.5
	NAM-1	75.4	61.1	48.7	62.1
	+	0.7	-0.4	-0.2	0.6
	NAM-6	75.3	61.4	48.9	61.9
	+	0.6	-0.1	0	0.4

TABLE 3. Performance comparison with other methods on NYUDv2-40 dataset.

Method	Pixel Acc	Mean Acc	Mean IoU	Fw IoU
FCN [6]	0.654	0.461	0.34	0.495
LSD-GF [43]	0.719	0.607	0.459	0.593
MMAF-Net [29]	72.2	0.592	0.448	-
RedNet [35]	0.813	0.603	0.478	-
RDF-152 [44]	0.815	0.601	0.477	-
CFN-152 [45]	0.813	0.603	0.481	-
TupleInfoNCE [46]	-	-	0.481	-
ACNet [13]	-	-	0.483	-
D-CNN [30]	-	0.611	0.484	-
MKE [47]	-	-	0.488	-
OURS	0.753	0.614	0.489	0.619

TABLE 4. Performance comparison on different architecture on the NYUDv2-40 dataset.

Archi-tecture	Settings	Pixel Acc	Mean Acc	Mean IoU	Fw IoU
Deep labV3+	Baseline	73.4	58.9	45.9	59.7
	NAM-6	74.4	59.7	47.3	60.8
	+	1	0.8	1.4	1.1
Deep labV3	Baseline	73.3	57.3	45.1	59.2
	NAM-6	74.7	60.2	47.5	61.1
	+	1.4	2.9	2.4	1.9
UNet	Baseline	70.9	54.7	42.1	57.7
	NAM-6	72.8	56.9	44.2	59.1
	+	1.9	2.2	2.1	1.4
PSPNet	Baseline	72.8	56.8	44.2	58.9
	NAM-6	70.7	53.5	41.3	56.5
	+	-2.1	-3.3	-2.9	-2.4
FPN	Baseline	72.8	57.3	44.7	59.1
	NAM-6	73.5	57.9	45.3	59.9
	+	0.7	0.6	0.6	0.8

TABLE 5. Ablation study of the proposed strategy on the NYUDv2-40 dataset.

Settings	Pixel Acc	Mean Acc	Mean IoU	Fw IoU
a. RGB	0.718	0.569	0.439	0.573
b. RGB+Depth	0.728	0.589	0.449	0.577
c. RGB+HHA	0.734	0.589	0.459	0.597
d. RGB + Depth + NAM-1	0.735	0.583	0.455	0.596
e. RGB + HHA + NAM-1	0.75	0.607	0.479	0.615
f. RGB + Depth + NAM-6	0.731	0.573	0.442	0.594
g. RGB + HHA + NAM-6	0.744	0.597	0.473	0.608

input them into the semantic segmentation network. The six layers consist of the 10th, the 20th, the 30th, the 40th, the 50th and the 60th layer. It can be seen from Tab.1. that our strategy outperforms the baseline methods in general on different backbones.

The results of the baseline methods and our method using different backbones on NYUDv2-40 are shown in Tab.2. In this table, the meaning of NAM-6 is the same as in Tab. 1, while NAM-1 means that we have performed channel-level concatenation of the 60th layer NAMLab hierarchical features, the RGB images, and the HHA depth images, which are taken as the input of the semantic segmentation network. It can be seen that our strategy achieves some improvement in general.

Conducting the experiments on the same NYUDv2-40 dataset, without modification or augmentation, it can be seen from Tab.3 that our strategy achieves better results for all four metrics in general.

B. EXPERIMENTS ON DIFFERENT ARCHITECTURES

Our proposed strategy is applied to the input stage of a semantic segmentation network, so it can be easily applied to most convolutional neural networks. Our method is also evaluated against a few representative semantic segmentation architectures such as DeepLabV3+, DeepLabV3, UNet, PSPNet, and FPN. The experimental results are shown in Tab.4 in order to determine whether it is generalizable or not.

It can be seen from Tab.4 that our strategy also achieved performance improvements on all architectures except for PSPNet.

C. ABLATION EXPERIMENTS

We have conducted ablation experiments to verify the effects of NAM-6 and NAM-1 features when stitching with different Depth and HHA images. As it can be seen from Tab.5, when the architecture is DeepLabV3+ and the backbone is ResNet101, both NAM-1 and NAM-6 settings can improve the performance of the semantic segmentation tasks no matter the Depth images or the HHA images are concatenated as the inputs. The RGB, the Depth and the HHA presented as the sittings in Tab.5 represents the feature type added to the input of the network.

V. CONCLUSION

There is now broad agreement that the performance of deep learning based segmentation algorithms is plateauing, especially in certain application domains such as medical image analysis. To advance to the next level of performance, we have further explored the combination of the CNN-based image segmentation models with the prominent "classical" NAMLab-based image segmentation method which was published recently. The core idea of the proposed algorithm is to take the channel-wise concatenation of the pre-segmentation labels provided by the traditional hierarchical image segmentation and the RGB-D features as the input of the neural network so as to guide the semantic segmentation tasks. In this paper, extensive experiments are conducted on the popular indoor RGB-D semantic segmentation datasets. When compared with the state-of-art algorithms, the experimental results presented in this paper show that our proposed method improves the performance of the image semantic segmentation networks on several popular neural network architectures.

However, there is still room for further improvement of the performance of the RGB-D indoor semantic segmentation model. In the future, we plan to design a unique and effective network architecture for extracting complementary information among the NAM layered features, or complementary information of the NAM features, the RGB features, and the Depth features. We believe that this work will further optimize the model for semantic segmentation tasks.

REFERENCES

[1] W. Zhou, J. Liu, J. Lei, L. Yu, and J.-N. Hwang, "GMNet: Graded-feature multilabel-learning network for RGB-thermal urban scene semantic segmentation," *IEEE Trans. Image Process.*, vol. 30, pp. 7790–7802, 2021.

[2] X. Ren, S. Ahmad, L. Zhang, L. Xiang, D. Nie, F. Yang, Q. Wang, and D. Shen, "Task decomposition and synchronization for semantic biomedical image segmentation," *IEEE Trans. Image Process.*, vol. 29, pp. 7497–7510, 2020.

[3] Y. Cai, L. Dai, H. Wang, and Z. Li, "Multi-target pan-class intrinsic relevance driven model for improving semantic segmentation in autonomous driving," *IEEE Trans. Image Process.*, vol. 30, pp. 9069–9084, 2021.

[4] S. Zhou, D. Nie, E. Adeli, J. Yin, J. Lian, and D. Shen, "High-resolution encoder–decoder networks for low-contrast medical image segmentation," *IEEE Trans. Image Process.*, vol. 29, pp. 461–475, 2020.

[5] T. Wu, S. Tang, R. Zhang, J. Cao, and Y. Zhang, "CGNet: A light-weight context guided network for semantic segmentation," *IEEE Trans. Image Process.*, vol. 30, pp. 1169–1179, 2021.

[6] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, Apr. 2017.

[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 2, pp. 84–90, Jun. 2012.

[8] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[10] Z. Zhang, "Microsoft Kinect sensor and its effect," *IEEE MultiMedia*, vol. 19, no. 2, pp. 4–10, Feb. 2012.

[11] Y. He, W.-C. Chiu, M. Keuper, and M. Fritz, "STD2P: RGBD semantic segmentation using spatio-temporal data-driven pooling," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7158–7167.

[12] X. Gao, M. Cai, and J. Li, "Improved RGBD semantic segmentation using multi-scale features," in *Proc. Chin. Control Decis. Conf. (CCDC)*, Jun. 2018, pp. 3531–3536.

[13] X. Hu, K. Yang, L. Fei, and K. Wang, "ACNET: Attention based network to exploit complementary features for RGBD semantic segmentation," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 1440–1444.

[14] Y. Zheng, B. Yang, and M. Sarem, "Hierarchical image segmentation based on nonsymmetry and anti-packing pattern representation model," *IEEE Trans. Image Process.*, vol. 30, pp. 2408–2421, 2021.

[15] N. Silberman, D. Hoiem, P. Kohli, R. Fergus, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, "Indoor segmentation and support inference from RGBD images," in *Computer Vision—ECCV*. Berlin, Germany: Springer, 2012, pp. 746–760.

[16] W. Liu, A. Rabinovich, and A. C. Berg, "ParseNet: Looking wider to see better," 2015, *arXiv:1506.04579*.

[17] O. Ronneberger, P. Fischer, T. Brox, J. Hornegger, W. M. Wells, and A. F. Frangi, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI*. Cham, Switzerland: Springer, 2015, pp. 234–241.

[18] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6230–6239.

[19] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected CRFs," 2014, *arXiv:1412.7062*.

[20] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Jun. 2016.

[21] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*.

[22] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder–decoder with atrous separable convolution for semantic image segmentation," 2018, *arXiv:1802.02611*.

[23] S. Minaee, Y. Y. Boykov, F. Porikli, A. J. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image segmentation using deep learning: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 7, pp. 3523–3542, Jul. 2022.

[24] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 5, pp. 603–619, May 2002.

[25] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient graph-based image segmentation," *Int. J. Comput. Vis.*, vol. 59, no. 2, pp. 167–181, Sep. 2004.

[26] T. Cour, F. Benezit, and J. Shi, "Spectral segmentation with multiscale graph decomposition," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 2, Jun. 2005, pp. 1124–1131.

[27] J.-H. Syu, S.-J. Wang, and L.-C. Wang, "Hierarchical image segmentation based on iterative contraction and merging," *IEEE Trans. Image Process.*, vol. 26, no. 5, pp. 2246–2260, May 2017.

[28] P. Arbeláez, M. Maire, C. Fowlkes, and J. Malik, "Contour detection and hierarchical image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 5, pp. 898–916, Aug. 2011.

[29] F. Fooladgar and S. Kasaei, "Multi-modal attention-based fusion model for semantic segmentation of RGB-depth images," 2019, *arXiv:1912.11691*.

[30] W. Wang and U. Neumann, "Depth-aware CNN for RGB-D segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 135–150.

[31] Y. Xing, J. Wang, and G. Zeng, "Malleable 2.5 D convolution: Learning receptive fields along the depth-axis for RGB-D scene parsing," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 555–571.

[32] C. Couprie, C. Farabet, L. Najman, and Y. LeCun, "Indoor semantic segmentation using depth information," 2013, *arXiv:1301.3572*.

[33] S. Gupta, R. Girshick, P. Arbeláez, and J. Malik, "Learning rich features from RGB-D images for object detection and segmentation," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 345–360.

[34] C. Hazirbas, L. Ma, C. Domokos, and D. Cremers, "FuseNet: Incorporating depth into semantic segmentation via fusion-based CNN architecture," in *Computer Vision—ACCV*. Cham, Switzerland: Springer, 2016, pp. 213–228.

[35] J. Jiang, L. Zheng, F. Luo, and Z. Zhang, "RedNet: Residual encoder–decoder network for indoor RGB-D semantic segmentation," 2018, *arXiv:1806.01054*.

[36] L.-Z. Chen, Z. Lin, Z. Wang, Y.-L. Yang, and M.-M. Cheng, "Spatial information guided convolution for real-time RGBD semantic segmentation," *IEEE Trans. Image Process.*, vol. 30, pp. 2313–2324, 2021.

[37] J. Cao, H. Leng, D. Lischinski, D. Cohen-Or, C. Tu, and Y. Li, "ShapeConv: Shape-aware convolutional layer for indoor RGB-D semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 7068–7077.

[38] H. Liang, S. Zhao, C. Chen, and M. Sarem, "The NAMlet transform: A novel image sparse representation method based on non-symmetry and anti-packing model," *Signal Process.*, vol. 137, pp. 251–263, Aug. 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0165168417300282

[39] Y. Zheng and M. Sarem, "A fast region segmentation algorithm on compressed gray images using non-symmetry and anti-packing model and extended shading representation," *J. Vis. Commun. Image Represent.*, vol. 34, pp. 153–166, Jan. 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1047320315002205

[40] J. Chen, S. Shan, C. He, G. Zhao, M. Pietikäinen, X. Chen, and W. Gao, "WLD: A robust local image descriptor," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1705–1720, Sep. 2010.

[41] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, and J. Garcia-Rodriguez, "A review on deep learning techniques applied to semantic segmentation," 2017, *arXiv:1704.06857*.

[42] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5987–5995.

[43] Y. Cheng, R. Cai, Z. Li, X. Zhao, and K. Huang, "Locality-sensitive deconvolution networks with gated fusion for RGB-D indoor semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1475–1483.

[44] S. Lee, S.-J. Park, and K.-S. Hong, "RDFNet: RGB-D multi-level residual feature fusion for indoor semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4990–4999.

[45] D. Lin, G. Chen, D. Cohen-Or, P.-A. Heng, and H. Huang, "Cascaded feature network for semantic segmentation of RGB-D images," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1320–1328.

[46] Y. Liu, Q. Fan, S. Zhang, H. Dong, T. Funkhouser, and L. Yi, "Contrastive multimodal fusion with TupleInfoNCE," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 754–763.

[47] Z. Xue, S. Ren, Z. Gao, and H. Zhao, "Multimodal knowledge expansion," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 854–863.

YUNPING ZHENG was born in Hubei, China, in 1979. He received the B.S. degree in computer science from the Air Force Early Warning Academy of the People's Liberation Army, Wuhan, Hubei, in 2001, and the M.S. and Ph.D. degrees in computer science from the Huazhong University of Science and Technology, Wuhan, in 2005 and 2008, respectively.

He was a Visiting Scholar with the Medical Image Processing Group, Department of Radiology, Perelman School of Medicine, University of Pennsylvania, from 2015 to 2016. He is currently an Associate Professor with the School of Computer Science and Engineering, South China University of Technology, Guangzhou, China. He has published more than 100 papers in refereed journals and conferences. His research interests include medical image processing, image representation, image compression, image segmentation, pattern recognition, and deep learning. He is a member of ACM and CCF.

YUAN XU was born in Guangdong, China, in 2000. He is currently pursuing the bachelor's degree in computer science and technology with the South China University of Technology.

His research interests include image processing and computer vision.

SHENGJIE QIU was born in Guizhou, China, in 2000. He is currently pursuing the bachelor's degree with the School of Computer Science and Engineering, South China University of Technology.

His research interests include image processing, deep learning, and natural language processing.

WENQIANG LI was born in Hunan, China, in 2001. He is currently pursuing the bachelor's degree with the School of Computer Science and Engineering, South China University of Technology, Guangdong.

His research interests include image segmentation, computer vision, and graph neural networks.

GUICHUANG ZHONG was born in Guangdong, China, in 2000. He is currently pursuing the bachelor's degree with the School of Computer and Engineering, South China University of Technology.

His research interests include computer vision and data mining.

MUDAR SAREM was born in Lattakia, Syria, in 1966. He received the B.S. degree in electronic engineering from Tishreen University, Lattakia, in 1989, and the M.S. and Ph.D. degrees in computer science from the Huazhong University of Science and Technology, Wuhan, China, in 1997 and 2002, respectively.

He was an Associate Professor with the School of Software Engineering, Huazhong University of Science and Technology, from 2005 to 2009, and a Visiting Professor, from 2010 to 2018. He is currently a main Researcher with the General Organization of Remote Sensing (GORS), Damascus, Syria. He has published more than 80 papers in refereed conferences and journals. His research interests include image processing, computer networks, and distributed systems.