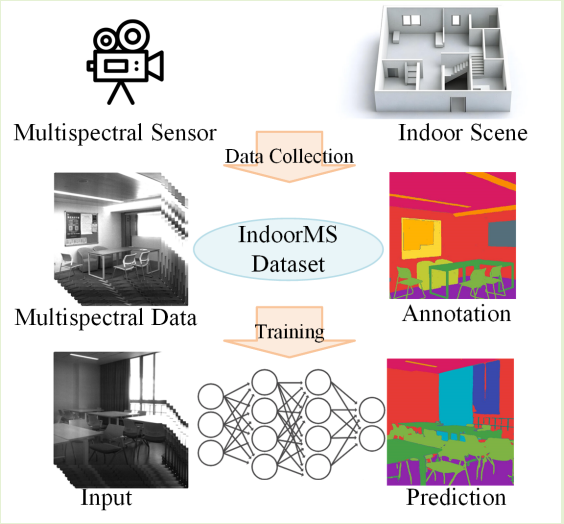
IndoorMS: A Multispectral Dataset for Semantic Segmentation in Indoor Scene Understanding

Qinfeng Zhu, *Graduate Student Member, IEEE*, Jingjing Xiao, and Lei Fan, *Senior Member, IEEE*

***Abstract*—Indoor scene understanding is a critical task in computer vision, traditionally relying on RGB data for deep learning-based semantic segmentation to achieve pixel-level understanding. However, indoor environments provide valuable information beyond the visible light spectrum, which has been largely overlooked in existing research. To address this gap, we introduce IndoorMS, a comprehensive multispectral dataset specifically designed for the semantic segmentation of indoor scenes. The dataset comprises images captured using a multispectral sensor in 17 buildings across diverse indoor settings, including meeting rooms, halls, lounges, offices, corridors, and classrooms. With 19 finely annotated semantic categories, IndoorMS enables robust evaluation of indoor scene segmentation. Benchmark experiments are performed using several leading semantic segmentation frameworks, followed by a thorough analysis of their performance. The results indicate that the optimal model combination, namely ConvNeXt-s with UperNet, achieved an mF1 score of 82.38 and an mIoU score of 72.90. Despite these promising results, IndoorMS's challenges on segmentation networks remain, such as class distribution imbalance and domain gaps between RGB and multispectral data. This work marks the first effort to support multispectral indoor scene understanding with a dedicated dataset, offering new opportunities for research in this domain. Potential avenues for future research are presented. The project page for the IndoorMS dataset is available at https://zhuqinfeng1999.github.io/IndoorMS/. (The dataset will be publicly available for download after peer review.)**



***Index Terms*—Multispectral, Image, Dataset, Semantic Segmentation, Indoor, Scene Understanding**

I. Introduction

Indoor scene understanding plays a crucial role in indoor intelligent and automated systems [1, 2]. Unmanned systems, such as indoor autonomous vehicles and drones, require an accurate understanding of their surroundings to identify various indoor objects and regions (e.g., doors, windows, walls, ceilings) in order to navigate, plan paths, and execute other complex tasks [3]. Moreover, indoor scene understanding [4] has potential applications in smart homes, intelligent security systems, and virtual and augmented reality environments [5].

Semantic segmentation of images is a crucial method for indoor scene understanding [6]. It involves assigning a category label to each pixel, enabling pixel-level understanding of the image [7]. In indoor scene understanding, semantic segmentation allows a computer to achieve fine-grained comprehension of indoor images, accurately segmenting various indoor elements into specific categories, such as walls, floors, doors, and windows.

In recent years, deep learning has become the dominant method for semantic segmentation [7, 8], largely due to the development of several highly effective deep neural networks. Networks like Fully Convolutional Network (FCN) [9], U-Net [10], the DeepLab series [11], as well as the more recent Vision Transformer (ViT) [12], and Vision Mamba [13, 14] have all demonstrated outstanding performance in semantic segmentation tasks. FCN replaces traditional fully connected layers with fully convolutional layers to achieve pixel-level mapping between input images and output feature maps [9]. U-Net utilizes a symmetric design of downsampling and upsampling, allowing the network to effectively combine multi-scale information [10], while DeepLab incorporates atrous convolution and Conditional Random Fields (CRF) to improve segmentation accuracy along boundaries [11]. The recently proposed ViT has a global receptive field, enabling it to capture contextual information effectively, making it particularly well-suited for complex scenes [12]. Vision Mamba, on the other hand, combines global receptive fields with linear complexity, showing great potential in handling high-resolution images [15].

This work was supported in part by the Xi'an Jiaotong-Liverpool University Research Enhancement Fund under Grant REF-21-01-003, and in part by the Xi'an Jiaotong-Liverpool University Postgraduate Research Scholarship under Grant FOS2210JJ03. (Corresponding author: Lei Fan.)

Qinfeng Zhu, and Lei Fan are with the Department of Civil Engineering, Xi'an Jiaotong-Liverpool University, Suzhou, 215123, China. (e-mail: Qinfeng.Zhu21@student.xjtlu.edu.cn; Lei.Fan@xjtlu.edu.cn)

Jingjing Xiao is with the Army Medical University, Chongqing, 400038, China. (email: shine636363@sina.com)

These deep learning networks rely on appropriate training data. In the field of semantic segmentation of indoor scenes, most existing research primarily uses RGB data for training. However, multispectral data have been largely overlooked for indoor scene understanding, despite their proven value across a wide range of applications, including remote sensing [16-18], medical imaging [19], mining [20], and cultural heritage preservation [21]. For instance, in agricultural remote sensing, multispectral data are used for effective monitoring of pests and diseases [22]; in forestry, they are utilized for tree species classification [23]; and in medicine, multispectral imaging is employed for skin disease diagnosis [19]. In these applications, multispectral data consistently offer more comprehensive features compared to RGB images. In the case of indoor scene understanding, multispectral data can also complement the RGB information, enabling computers to perceive additional dimensions of indoor scene properties.

Multispectral data are acquired by specific multispectral sensors that are capable of capturing the reflectance information of real-world objects across multiple distinct spectral bands [24]. This is achieved by utilizing multiple filters or different sensor modules to separate specific spectral bands, resulting in images that contain multiple spectral channels [25]. Compared to RGB sensors, multispectral sensors offer the advantage of capturing information beyond the visible spectrum, which typically includes both visible light bands as well as several infrared bands. Additionally, Glatt et al. [26] proposed a novel algorithm for illumination spectral estimation. Their study involved the collection of multispectral data from both indoor and outdoor environments. However, it did not include image analysis tasks such as object detection or segmentation. Their work [26] has significantly advanced the integration of multispectral cameras into mobile devices, and demonstrated the potential of multispectral imaging to enhance indoor scene understanding.

Despite the significant progress made in utilizing multispectral information across various application domains, its use in indoor environments remains relatively underexplored. To promote its application, we introduce IndoorMS, a multispectral dataset specifically designed for indoor scene understanding. Three example images from the IndoorMS dataset are shown in Fig. 1. These images were acquired using a multispectral sensor in challenging and complex indoor environments across 17 buildings. To account for variations in natural light, data collection was conducted under a diverse range of weather conditions and times of day. For IndoorMS, we developed a highly detailed set of semantic labels comprising 19 categories and implemented a rigorous annotation process. To ensure accuracy, we opted against the use of AI-assisted annotation, relying instead on fully manual labeling. We also evaluated the dataset using a variety of representative semantic segmentation frameworks to provide benchmark segmentation performance.

The main contributions of this work are as follows:

1. We present IndoorMS, the first multispectral dataset specifically designed for indoor scene understanding, providing essential data support for research on indoor scene understanding based on multispectral information.
2. We provide highly detailed semantic annotations for IndoorMS, making the segmentation tasks more challenging and valuable for future research.
3. We establish benchmark segmentation performance using a set of representative semantic segmentation methods and provide discussions on the challenges identified in the dataset.

The remainder of this paper is organized as follows: Section II provides an overview of related work. Section III presents a detailed description of the IndoorMS dataset. Section IV establishes benchmark performance using representative semantic segmentation methods. Section V discusses the challenges associated with indoor multispectral scene understanding. Finally, Section VI proposes future research directions and concludes the paper.

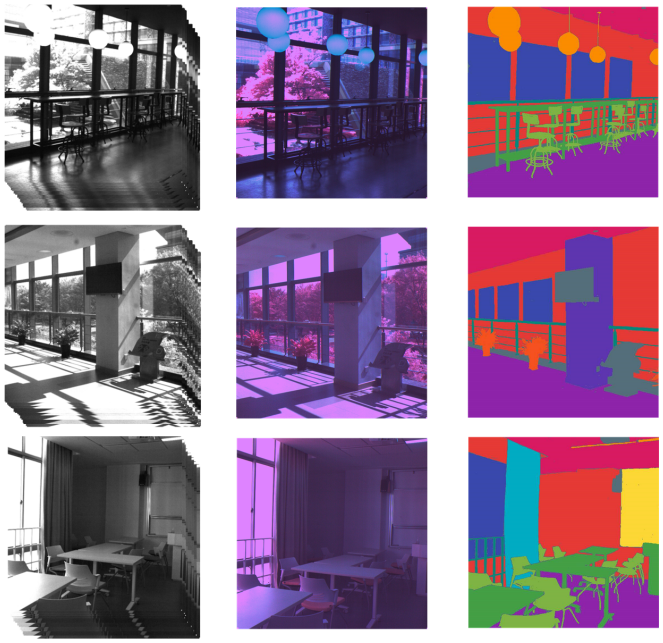


Fig. 1. Example images of the IndoorMS dataset, including a lounge scene (the first row), a corridor scene (the second row), and a classroom scene (the third row).

II. Related Work

To date, numerous datasets have been developed for indoor scene understanding, with the primary sensing modality still relying on RGB information [6]. The ADE20K dataset [27] includes a large number of indoor scenes with comprehensive annotations, making it one of the most widely used image segmentation datasets for indoor scene understanding. In addition, many datasets have also incorporated depth information, using RGB-D sensors for data collection to meet the requirements of 3D scene reconstruction tasks [28]. For example, Silberman et al. [29] collected 1,449 indoor RGB-D images, while ScanNet [30] collected 1,500 scanned scenes with 2.5 million frames of RGB-D images.

The annotation of real-world indoor scene image datasets is costly, prompting interest in synthetic datasets that eliminate the need for annotation [31]. Studies such as InteriorNet [32] and SceneNet [33] have synthesized large numbers of indoor scenes, simulating the image capture process to generate RGB

images, IMU data, and depth information. Their studies have demonstrated the effectiveness of employing synthetic data for pre-training a model that can improve its performance on real-world image data in vision tasks. However, generating synthetic multispectral data is challenging, primarily because accurately simulating the material properties of objects in non-visible spectral bands is difficult due to the complex interactions between materials and light beyond the visible spectrum. Current generation techniques, such as GAN [34] and diffusion models [35], often fail to accurately replicate objects' multi-spectral information in real-world scenarios [36], limiting the applicability of synthetic multispectral data in training robust deep learning models.

Unlike synthetic datasets, real-world multispectral data realistically capture the spectral characteristics of indoor environments, which are valuable for applications such as material recognition and scene understanding under diverse lighting conditions. Although significant progress has been made in leveraging RGB and depth information for real-world indoor scene understanding, the exploration of non-visible spectral information for this task remains underexplored. Therefore, these justify the necessity of our work in collecting real-world multispectral data and providing fine semantic labels to establish a real-world multispectral dataset, named IndoorMS, for indoor scene understanding.

III. The IndoorMS Dataset

A. Data Collection

The Silios CMS4 multispectral sensor is used in this study for multispectral data collection. This sensor utilizes a 2048 × 2048 resolution CMOS imaging chip, equipped with a miniature interferometric filter array to achieve multispectral sampling. The filter array is arranged in a 3 × 3 mosaic pattern, allowing for the acquisition of 8 multispectral channels and 1 grayscale channel within each "super-pixel," with a resulting resolution of 682 × 682 for these channels. The details of these 9 channels are presented in TABLE I.

TABLE I
Spectral Band Characteristics of the CMS4 Multispectral Sensor

Band	λ_C (nm)	FWHM (nm)	T_{max} (%)
1	554	38	52
2	591	36	53
3	628	34	54
4	667	33	54
5	719	33	52
6	758	32	50
7	797	32	48
8	838	33	45
9	Neutral Density	-	T_{mean} = 12% over 500–900 nm

λ_C = central wavelength, FWHM = the full width at half maximum, T_{max} = the maximum transmittance, T_{mean} = the mean transmittance.

To ensure maximum diversity of indoor scenes, we collected data from 17 different buildings, including educational buildings, research office buildings, and public service facilities, encompassing a wide range of architectural functions and layouts. The data collection covered various indoor scenes such as meeting rooms, halls, lounges, offices, corridors, and classrooms, among others.

Considering that lighting conditions can significantly impact the richness of information captured across different spectral bands in indoor environments [37], we paid particular attention to capturing diverse lighting conditions during the data collection process, including combinations of natural and artificial indoor lighting. Therefore, we conducted data collection at different times of the day and under various weather conditions to encompass a broad range of lighting scenarios. Our data collection spanned from 9 AM to 5 PM, effectively covering the lighting changes from morning to late afternoon. The weather conditions included sunny, cloudy, overcast, and rainy days, allowing us to capture indoor environments under a variety of natural lighting conditions.

In different indoor scenes, the distribution of light sources also varied. For instance, some scenes relied heavily on abundant natural light, often due to the presence of large windows, while others were primarily illuminated by indoor artificial lighting, especially in spaces located deeper within buildings or on lower floors. Additionally, we specifically collected data in low-light conditions, where the available illumination was minimal.

We also ensured diversity in camera perspectives during data collection. In each scene, we captured images from multiple different viewpoints and heights to capture the varying spatial and object features present in the scene. Furthermore, we aimed to avoid redundant data collection of the same scene to ensure that each image represented unique lighting characteristics and spatial configurations.

TABLE II
Semantic Segmentation Categories and Descriptions

Category	Description
Clutter	Items that do not belong to the defined categories below.
Wall	Walls of various materials, including glass walls.
Ceiling	Ceilings, excluding items such as lights or surveillance cameras.
Floor	The ground surface of the indoor environment.
Column	Freestanding columns that are not adjacent to any walls on all four sides.
Window	Windows of various materials, including their attachments (e.g., frames).
Door	Doors of various materials, including glass doors and their attachments (e.g., handles, frames).
Elevator	Elevator entrances, including their attachments.
Curtain	Curtains used for windows.
Railing	Railings used for staircases or balconies.
Table	Tables of various sizes and shapes, including desks and dining tables.
Chair	Chairs of various types, including office chairs and stools.
Sofa	Sofas, typically used in lounges or waiting areas.
Board	Blackboards and whiteboards.
Poster	Posters and their attachments (e.g., frames).
Light	Light fixtures, including ceiling lights and wall-mounted lamps.
Plant	Indoor plants, including their pots and other attachments.
Bin	Waste bins of various sizes and types.

B. Data Annotation

In defining the semantic segmentation categories for IndoorMS, we considered both the representativeness of semantic categories and their practical applicability, covering a

range of semantic objects from structural elements to functional furniture, and down to finer details. Specifically, we defined 19 categories in our dataset, and their detailed descriptions can be found in TABLE II. This comprehensive categorization provides robust data support for various indoor intelligent applications. We selected important and common structural elements and furniture found in indoor scenes (such as walls, ceilings, floors, doors, and windows), which are critical for scene understanding and 3D reconstruction. By perceiving these categories, a semantic segmentation model can achieve an understanding of the overall room structure, providing a foundation for subsequent tasks such as navigation and path planning [38, 39].

Additionally, our defined categories include those essential for human-computer interaction and indoor activities (such as tables, chairs, sofas, boards, and posters). These objects reflect different usage scenarios, such as offices, classrooms, and lounges. Understanding these categories helps in identifying the functional layout of different scenes and provides targeted training data for service robots. Furthermore, we considered specific detail-oriented targets that are unique to indoor scenes (such as plants and waste bins). These categories often appear in complex scenes and are important for applications such as obstacle avoidance in robotics.

Since the collected multispectral data contains information from 9 spectral bands, directly displaying grayscale images of these bands is not conducive to annotation. Therefore, prior to annotating the multispectral data, it is necessary to visualize the information from all 9 channels, allowing annotators to have an intuitive understanding of the spectral characteristics of each pixel. To achieve this, we employed pseudo-color conversion techniques to transform the multispectral images into a more visually intuitive RGB format, thereby better presenting the scene features [40].

Specifically, we defined a pseudo-color conversion matrix M to map the information from the 9 spectral bands to the RGB channels. Let I represent the vector composed of the 9 spectral bands of the multispectral data, and O represent the resulting RGB vector after pseudo-color conversion. The pseudo-color conversion process can be represented by Eq. (1):

$$\mathrm{O} = \mathrm{M} \cdot \mathrm{I} \tag{1}$$

Here, $\mathrm{I} = [I_1, I_2, I_3, \ldots, I_9]^T$ represents the input signals from the 9 spectral bands, and $\mathrm{O} = [R, G, B]^T$ represents the RGB output signals after pseudo-color conversion. The matrix M is the pseudo-color conversion matrix used for the transformation, and its values are given by Eq. (2):

$$\mathrm{M} = \begin{bmatrix} 0.191 & 0.001 & -0.401 & -0.545 & 0.789 & 1.048 & 0.131 & -0.132 & -0.212 \\ -0.162 & -0.135 & 0.410 & 0.514 & 0.096 & -0.123 & -0.067 & -0.005 & -0.015 \\ 1.038 & 0.411 & -0.032 & 0.035 & 0.043 & 0.099 & 0.118 & 0.057 & -0.769 \end{bmatrix} \tag{2}$$

In matrix M, each row represents the weighting coefficients used to map the 9 spectral band signals to the RGB channels. The first row is used to calculate the red channel, the second row for the green channel, and the third row for the blue channel. Through this mapping, the multispectral information can be presented in the form of a pseudo-color RGB image, as illustrated in Fig. 2.

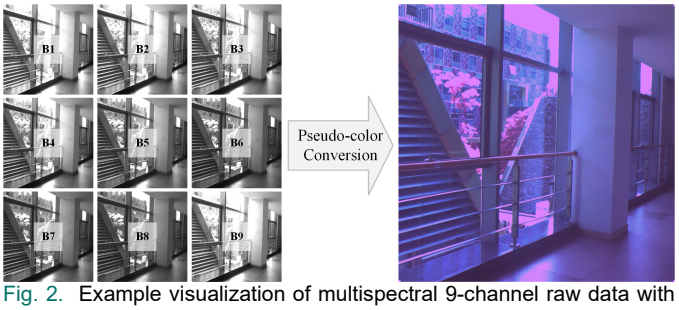


Fig. 2. Example visualization of multispectral 9-channel raw data with the pseudo-color conversion.

To ensure the accuracy of annotations, this dataset was uniformly annotated without using AI-assisted tools, such as Segment Anything Model (SAM) [41] or other automatic segmentation techniques. Instead, we employed a fully manual annotation approach, with Label Studio being used as the annotation tool. In the annotation process, pseudo-color images were utilized for annotation purposes. To avoid potential annotation errors caused by discrepancies between the pseudo-color images and real-world colors, we also provided high-resolution color images captured by a camera for reference. This allowed annotators to compare the pseudo-color images with the real color images, thus minimizing errors.

A schematic of the annotation process is illustrated in Fig. 3. Fig. 3(a) shows the pseudo-color image, on which different semantic categories were manually annotated using the annotation software. The resulting annotated image is shown in Fig. 3(b). During the annotation process, we provided high-resolution RGB images captured by the camera, as illustrated in Fig. 3(c). Finally, the resulting ground truth of segmentation is presented in Fig. 3(d).

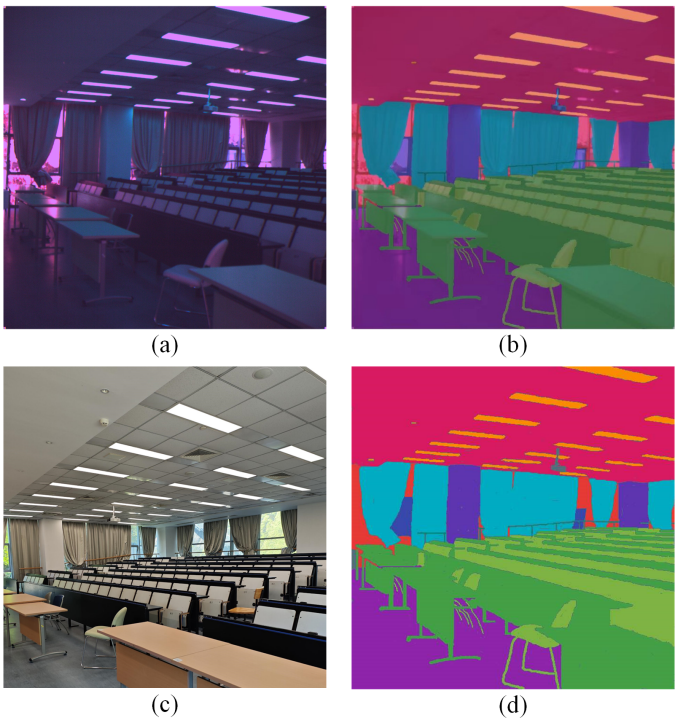


Fig. 3. Example of the annotation process. (a) Pseudo-color image for annotation; (b) Annotated pseudo-color image; (c) High-resolution color image provided for reference during annotation; (d) Resulting ground truth of segmentation.

We recruited experienced individuals for the annotation task, dividing them into annotation and validation teams. To ensure consistency in annotations, all annotators and validators underwent unified training. The annotation team was responsible for the initial annotations, while the validation team checked the results. Any annotations that did not meet the required standards were returned to the annotation team for correction until all annotations were accurate and complete.

C. Dataset Split

The dataset consists of a total of 227 multispectral images, which were divided into training, validation, and test sets in a 7:1:2 ratio. Specifically, the training set contains 158 images, the validation set contains 22 images, and the test set contains 47 images. We employed a random splitting strategy to partition the dataset. This approach ensures that the training, validation, and test sets all retain scene diversity. Additionally, it helps ensure that rare categories (e.g., plants, bins) are balanced across both the training and testing phases.

IV. Benchmarks

A. Representative Baselines

To comprehensively evaluate the performance of the IndoorMS dataset in semantic segmentation tasks, we conducted benchmark testing using various representative semantic segmentation frameworks. Common semantic segmentation frameworks are typically composed of two main components: an encoder and a decoder [42]. The encoder is responsible for extracting multi-scale features from the input image, while the decoder maps these features back to pixel-level class labels to generate segmentation predictions. In this study, we selected multiple representative combinations of encoders and decoders to thoroughly demonstrate the performance of IndoorMS under different segmentation architectures.

ResNet [43] is a classic and widely-used convolutional neural network-based encoder architecture [44]. It introduces skip connections to address the vanishing gradient problem in deep networks, allowing the construction of deeper networks that are more effective in feature learning. ConvNeXt [45] simplifies the traditional convolutional network and incorporates Transformer-inspired [12] design elements. Its extensively optimized architecture allows for efficient feature extraction when dealing with complex structures and textures. In addition to convolutional neural network-based encoders, we also selected the Swin Transformer [46], a visual Transformer architecture based on a sliding window mechanism. Leveraging the self-attention mechanism, Swin Transformer is capable of capturing long-range dependencies in the image, providing efficient global feature modeling.

PSPNet [47] is a pyramid pooling decoder architecture that utilizes different pooling operations to capture the global contextual information of the image, enabling effective feature extraction from different scales of image regions. UperNet [48] is a multi-level decoder that integrates features from different levels of the encoder to effectively handle multi-scale information. BiSeNet [49] combines spatial and context paths to process features, with the spatial path retaining high-resolution spatial features and the context path capturing global semantic information. DeepLabV3+ [11] employs atrous convolution, which allows for a broader receptive field during feature extraction. We also selected the Transformer-based decoder SegFormer [50], which achieves end-to-end segmentation by integrating with a Transformer encoder and excels at fine-grained feature extraction.

B. Experimental Setup and Evaluation Metrics

In constructing the semantic segmentation framework, we adopted multiple encoder-decoder combinations for benchmark experiments, including ConvNeXt with UperNet, ResNet with BiSeNet, ResNet with DeepLabV3+, Mix ViT with SegFormer, ResNet with PSPNet, Swin Transformer with UperNet, and ResNet with UperNet. To evaluate the impact of network depth on segmentation performance, we selected two different scales of models for each combination. Specifically, for the encoders, we used ResNet18 and ResNet50, ConvNeXt's Tiny and Small versions, Swin Transformer's Tiny and Small versions, as well as Mix ViT's b0 and b2 versions.

To enhance the generalization ability of the tested models, we implemented a range of data augmentation strategies [51], including random resizing, random rotation, random flipping, and random cropping. All input images were resized to a resolution of 512 × 512, with a batch size of 16 (8 images per GPU). The models, with the exception of SegFormer that was trained for 160k iterations, were trained for 15k iterations. We used Cross-Entropy Loss as the loss function to optimize the models. It is important to note that the choice of these parameters was the result of extensive hyperparameter tuning and optimization to achieve the best performance across the different model combinations on the IndoorMS dataset. The final, optimized experimental settings are presented in TABLE III. All experiments were conducted using two 24GB NVIDIA 4090D GPUs.

TABLE III
Training Settings for Segmentation Framworks on the IndoorMS Dataset

Decoder	Encoder	LR	OP	SC	WA
UperNet	ConvNeXt	0.001	AdamW	PolyLR	Yes
BiSeNet	ResNet	0.025	SGD	PolyLR	Yes
DeepLabV3+	ResNet	0.1	SGD	PolyLR	No
Segformer	Mix ViT	0.0006	AdamW	PolyLR	Yes
PSPNet	ResNet	0.01	SGD	PolyLR	No
UperNet	Swin	0.00006	AdamW	PolyLR	Yes
UperNet	ResNet	0.1	SGD	PolyLR	No

LR: Learning Rate, OP: Optimizer, SC: Schedule, WA: Warmup, Swin: Swin Transformer.

To evaluate the performance of semantic segmentation networks on the IndoorMS dataset, we used three evaluation metrics: Intersection over Union (IoU), mean Intersection over Union ($mIoU$), and mean $F1$ score ($mF1$). Among these, $mIoU$ and $mF1$ are used to comprehensively evaluate the overall performance of the semantic segmentation models, while IoU is used to reflect models' performance across different categories in detail. To eliminate any potential experimental variability, we performed three complete training experiments for each encoder-decoder framework. The average of the metric results from these repeated trainings was used to represent the performance of each network framework. This

TABLE IV
THE BENCHMARK SEGMENTATION RESULTS OF REPRESENTATIVE METHODS WITHOUT PRETRAINING

Decoder	Encoder	mF1	mIoU	Clutter	Wall	Ceiling	Floor	Column	Window	Door	Elevator	Curtain	Railing	Table	Chair	Sofa	Board	Poster	Light	Plant	Bin	Signage
UperNet	ConvNeXt-t	65.50	51.61	21.63	74.26	78.21	83.11	30.89	63.87	37.77	27.48	64.02	47.42	54.95	55.41	85.24	41.18	44.54	65.00	46.24	50.59	8.76
	ConvNeXt-s	71.71	50.73	20.74	73.90	80.89	79.51	28.79	66.36	38.28	19.78	65.99	45.68	53.81	55.83	82.99	56.99	17.93	67.77	47.23	53.10	8.29
BiSeNet	ResNet18	53.01	39.60	16.52	67.22	59.88	78.83	19.81	46.62	32.65	28.52	56.72	22.85	45.59	44.38	80.93	26.45	13.85	59.87	37.31	9.73	4.64
	ResNet50	50.20	36.84	15.80	63.15	54.61	75.56	14.68	45.69	24.49	15.29	55.03	22.95	41.49	40.73	76.59	26.27	12.27	60.30	33.35	17.53	4.22
DeepLabV3+	ResNet18	75.62	42.89	19.92	67.56	63.30	83.26	14.35	52.83	37.63	15.97	67.97	34.89	43.30	50.84	83.46	32.10	15.91	61.75	41.03	21.16	7.75
	ResNet50	63.81	50.12	22.57	73.10	74.83	85.69	31.64	61.70	44.40	30.92	73.10	38.97	55.04	60.94	86.05	35.71	26.84	68.16	46.96	20.95	14.77
Segformer	Mix ViT-b0	71.49	50.16	21.33	71.57	80.38	81.81	29.81	54.38	46.05	33.13	65.33	41.54	54.35	52.34	82.53	37.99	45.43	68.13	48.74	27.13	11.02
	Mix ViT-b2	53.30	40.04	18.48	66.44	63.21	79.89	18.30	43.38	36.30	16.21	39.63	38.35	50.56	46.23	81.26	24.75	12.88	65.62	44.65	6.90	7.65
PSPNet	ResNet18	46.38	33.61	16.48	58.77	44.65	73.84	11.47	36.33	23.08	5.72	52.61	21.45	39.79	36.01	77.93	21.13	21.79	54.63	35.03	4.86	2.97
	ResNet50	45.17	32.89	13.70	59.78	42.09	74.92	9.75	36.16	24.14	4.87	60.09	19.92	35.67	32.98	76.30	20.13	16.28	57.35	34.08	3.97	2.77
UperNet	Swin-t	55.36	41.45	17.89	65.05	59.60	80.87	19.69	39.50	28.02	14.94	52.87	36.79	43.67	48.25	81.72	30.60	24.09	66.98	44.99	26.22	5.80
	Swin-s	57.20	43.32	18.23	66.55	66.04	81.08	18.99	42.79	31.71	15.42	61.74	36.37	44.87	49.27	82.43	31.34	34.51	66.96	42.36	25.59	6.78
UperNet	ResNet18	62.00	47.93	21.48	72.44	67.89	85.14	28.35	56.37	41.95	35.24	67.49	40.13	46.50	55.19	83.40	28.73	44.03	65.79	40.85	17.70	12.07
	ResNet50	45.17	32.89	13.70	59.78	42.09	74.92	9.75	36.16	24.14	4.87	60.09	19.92	35.67	32.98	76.30	20.13	16.28	57.35	34.08	3.97	2.77

TABLE V
THE BENCHMARK SEGMENTATION RESULTS OF REPRESENTATIVE METHODS WITH PRETRAINING

Decoder	Encoder	mF1	mIoU	Clutter	Wall	Ceiling	Floor	Column	Window	Door	Elevator	Curtain	Railing	Table	Chair	Sofa	Board	Poster	Light	Plant	Bin	Signage
UperNet	ConvNeXt-t	79.89	69.29	26.96	83.55	86.37	92.48	46.31	73.37	70.12	75.83	83.65	58.37	66.57	74.29	92.48	68.25	81.85	67.21	72.01	79.86	16.90
	ConvNeXt-s	82.38	72.90	28.20	84.64	88.53	92.65	56.20	73.93	68.88	85.77	90.86	57.70	68.82	75.01	92.49	86.60	88.28	65.96	72.44	88.48	19.63
BiSeNet	ResNet18	70.43	51.82	21.04	74.07	78.27	77.46	30.98	59.86	37.31	58.10	85.59	27.30	50.76	43.97	80.02	47.86	30.15	60.70	48.39	53.39	19.36
	ResNet50	70.68	57.65	21.70	76.58	81.30	84.71	36.34	66.30	44.74	60.08	87.75	28.95	54.47	60.02	85.09	47.42	73.15	61.63	51.90	57.89	15.32
DeepLabV3+	ResNet18	73.34	51.59	21.49	72.00	68.38	81.73	23.10	59.65	45.59	47.72	82.25	34.63	48.89	56.26	83.96	34.14	47.03	66.44	42.03	55.51	9.35
	ResNet50	69.31	56.12	23.44	75.56	82.61	85.53	29.14	65.68	47.86	55.08	87.26	36.48	52.95	65.09	85.31	41.67	46.55	65.27	45.38	63.01	12.37
Segformer	Mix ViT-b0	70.65	57.49	24.30	77.07	81.24	86.92	40.76	61.44	47.92	75.04	78.19	45.16	59.04	62.66	86.32	53.22	33.26	69.50	54.07	43.05	13.22
	Mix ViT-b2	77.39	65.90	26.21	81.45	85.97	90.55	52.03	67.62	65.63	69.65	85.49	50.27	63.74	68.05	87.05	76.27	79.36	71.88	59.20	58.66	13.11
PSPNet	ResNet18	57.75	43.95	17.22	69.21	70.15	81.49	25.52	49.11	34.70	34.74	74.77	24.76	41.02	51.45	80.90	29.92	9.38	56.75	38.63	32.06	13.24
	ResNet50	67.16	53.37	21.38	73.65	77.78	84.06	30.86	57.41	41.46	34.58	80.90	29.12	47.65	58.90	86.87	38.43	55.34	58.58	51.25	66.46	19.42
UperNet	Swin-t	70.13	57.42	23.81	76.68	77.95	88.79	46.38	65.14	53.41	46.86	81.31	44.68	57.47	63.81	86.37	57.25	15.17	63.73	63.05	68.58	10.52
	Swin-s	78.22	67.36	27.44	81.01	86.18	89.32	49.81	67.80	58.97	81.20	92.87	47.26	62.30	68.47	90.34	67.64	85.40	66.80	71.65	73.85	11.51
UperNet	ResNet18	76.67	56.39	22.75	74.63	78.26	85.48	29.37	64.40	44.93	57.37	82.04	42.54	54.10	57.16	82.55	49.85	69.12	66.85	41.22	57.39	11.40
	ResNet50	69.15	55.73	23.29	76.00	78.25	86.10	30.91	61.54	47.78	58.20	86.59	41.64	55.31	63.67	86.93	40.42	48.68	67.44	49.63	42.95	13.60

ensured our findings were more robust and less influenced by random fluctuations. The formula for IoU is given by Eq. (3):

$$IoU = \frac{TP}{TP + FP + FN} \tag{3}$$

Here, True Positive (TP) represents the number of pixels correctly classified as positive, False Positive (FP) represents the number of pixels incorrectly classified as positive, and False Negative (FN) represents the number of positive pixels incorrectly classified as negative. The higher the IoU value, the better the segmentation performance. Therefore, $mIoU$ can be used to evaluate the overall performance of semantic segmentation, which can be expressed by Eq. (4):

$$mIoU = \frac{1}{N}\sum_{i=1}^{N} IoU_i = \frac{1}{N}\sum_{i=1}^{N} \frac{TP_i}{TP_i + FP_i + FN_i} \tag{4}$$

Here, i represents the i-th class, and N represents the total number of classes. In addition to $mIoU$, which evaluates overall segmentation performance, $mF1$ is calculated as the average of the $F1$ scores across all classes. The $F1$ score is the harmonic mean of precision and recall, providing another measure of the model's overall performance, and can be expressed by Eq. (5):

$$mF1 = \frac{1}{N}\sum_{i=1}^{N} F1_i = \frac{1}{N}\sum_{i=1}^{N} \frac{2 \cdot TP_i}{2 \cdot TP_i + FP_i + FN_i} \tag{5}$$

C. Benchmark Results

We trained several representative networks on the IndoorMS dataset from scratch, and the results of semantic segmentation performance are summarized in TABLE IV. It is evident that training from scratch led to limited segmentation performance across all methods. Among these, the best-performing combination was ConvNeXt-s with UperNet, which achieved an $mF1$ score of 65.50 and an $mIoU$ score of 51.61. Furthermore, deeper networks did not always result in superior segmentation performance. For example, when using BiSeNet as the decoder, ResNet50 performed worse in terms of $mIoU$ compared to ResNet18. This can be attributed to the relatively small scale of the IndoorMS dataset, as well as the richer features inherent in multispectral data compared to conventional RGB images. Given that mainstream neural networks typically require large-scale datasets to perform optimally, training from scratch alone was insufficient for effectively fitting the training data.

In semantic segmentation tasks, to achieve better segmentation results, it is common to pre-train the encoder (i.e., the backbone of the segmentation framework) [52]. Specifically, the backbone network is first pre-trained on a large-scale image dataset (such as ImageNet [53]) for image classification, enabling the model to learn rich feature extraction capabilities beforehand. Through such pre-training, the encoder effectively captures features at various levels—from low-level to high-level—which can then be fine-tuned for the downstream segmentation task, significantly enhancing the segmentation performance. Therefore, in our experiments, we also pre-trained the encoder on ImageNet to assess the impact of pre-training on the IndoorMS dataset. The segmentation results after pre-training and fine-tuning are shown in TABLE V.

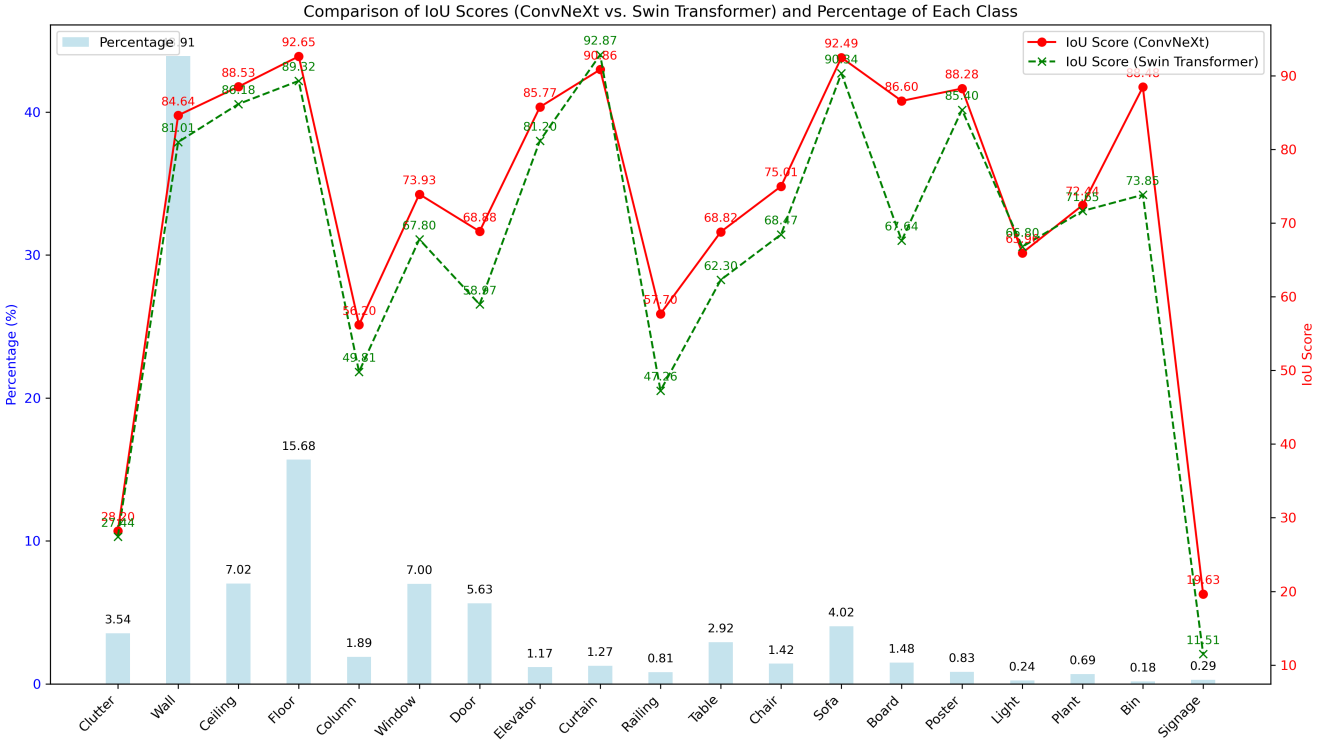


Fig. 4. The comparison of the pixel proportion of various categories in IndoorMS and the IoU scores of ConvNeXt and Swin Transformer's small version models.

Under the pre-training and fine-tuning paradigm, tested networks on the IndoorMS dataset showed significant performance improvements. The best-performing method was the combination of ConvNeXt-s and UperNet, achieving an $mF1$ score of 82.38 and an $mIoU$ score of 72.90. It is evident that incorporating multispectral data allowed state-of-the-art segmentation frameworks to achieve strong performance across most of the 19 defined categories, with IoU scores generally exceeding 70. Furthermore, when deeper networks were used after pre-training, they demonstrated better performance due to their enhanced ability to learn from the larger ImageNet dataset. As the depth of the network increased, so did its capacity to capture more complex features, leading to better segmentation results during the fine-tuning process. However, some specific categories still had relatively lower segmentation scores, including Clutter, Column, Railing, and Signage.

Using larger models comes at the cost of increased computational load and parameter size, which may be a consideration for practical applications. However, thanks to the rapid advancements in GPU technology, modern GPUs are now capable of efficiently running larger models. While the inference speed of larger models is slightly slower compared to smaller models, it remains fast enough for practical use. In our tests, using the NVIDIA 4090D GPU, the inference speed for 9-channel 512×512 resolution images was approximately 1.33ms per image when using the Tiny model (or ResNet-18) and 1.78ms per image when using the Small model (or ResNet-50). This efficiency is deemed more than sufficient for real-time segmentation applications.

The Clutter category had a low score because IndoorMS contains a large number of complex and challenging scenes, which include many objects that the model has not encountered before, making accurate segmentation difficult. The Column category also proved challenging because our definition of Column refers to pillars that are not adjacent to any walls, whereas in the dataset, some columns are adjacent to one wall, making it difficult for the model to distinguish based on image data alone. Railing and Signage had lower scores due to the rarity of these categories in the training data and their small size, making their prediction particularly challenging.

To further substantiate the advantages of multispectral data in semantic segmentation tasks, we conducted a series of comparative experiments that aimed at providing preliminary evidence of the potential performance enhancement offered by multispectral data over RGB data. Specifically, we performed semantic segmentation tests using pseudo-color images (simulating RGB data due to the spectral band limitation of the adopted sensor that is unable to capture all red, green and blue spectra) and multispectral data, respectively, and subsequently compared their segmentation performance results. The experimental outcomes are presented in TABLE VI. To ensure the fairness of the comparison, we employed models pre-trained on the ImageNet dataset for segmenting both multispectral and pseudo-color images. The results demonstrate that the semantic segmentation performance using multispectral data as the input significantly surpassed that achieved with the pseudo-color images as the input. However, it is important to note that this experiment serves only as an initial validation using pseudo-color images. To facilitate a more rigorous comparative analysis, it is necessary to employ multi-spectral data with a more comprehensive set of channels covering the red, green, and blue spectra.

TABLE VI
COMPARISON OF SEMANTIC SEGMENTATION PERFORMANCE OF MULTISPECTRAL DATA AND PSEUDO-COLOR IMAGES

Decoder	Encoder	Data	mIoU
UperNet	ConvNeXt-s	Multispectral	72.90
		Pseudo-color	70.61
BiseNet	ResNet50	Multispectral	57.65
		Pseudo-color	51.42
DeepLabV3+	ResNet50	Multispectral	56.12
		Pseudo-color	55.17
Segformer	Mix ViT-b2	Multispectral	65.90
		Pseudo-color	64.06
PSPNet	ResNet50	Multispectral	53.37
		Pseudo-color	50.94
UperNet	Swin-s	Multispectral	67.36
		Pseudo-color	65.21
UperNet	ResNet50	Multispectral	55.73
		Pseudo-color	49.15

D. Differences from other Datasets

The IndoorMS dataset differs significantly from existing datasets in several key aspects. First and foremost, IndoorMS is the first multispectral dataset specifically designed for indoor scene understanding, incorporating multiple spectral channels beyond the visible light spectrum. This is a distinctive feature that sets it apart from conventional datasets that primarily rely on RGB data [27]. The inclusion of multispectral data offers additional spectral dimensions of object information, enabling more comprehensive scene understanding.

Secondly, IndoorMS provides a finer level of semantic categorization, comprising 19 distinct categories. This level of granularity surpasses that of many existing datasets [6], which often feature fewer, broader categories. The detailed annotations in IndoorMS enhance its applicability for more complex semantic segmentation tasks, allowing for more precise and targeted model evaluations.

Finally, the complexity and diversity of the segmented scenes within IndoorMS further distinguish it from other datasets [32]. The diversity of indoor environments, coupled with the carefully crafted segmentation labels, provides a rich and challenging benchmark for evaluating the capabilities of semantic segmentation architectures. These characteristics make IndoorMS particularly valuable for testing models designed to handle complex, real-world indoor scene understanding tasks.

V. CHALLENGES

A. Class Distribution Imbalance

In indoor image data, there is a significant imbalance in class distribution [54]. This is because, during indoor scene capture, most of the image pixels are concentrated on common structures such as walls, ceilings, floors, doors, and windows, while other less common objects are comparatively rare. We analyzed the pixel distribution for each class in the IndoorMS dataset, as shown in the Fig. 4. The categories "Wall," "Floor," "Ceiling," and "Window" occupy substantial proportions, accounting for 43.91%, 15.68%, 7.02%, and 7.00%, respectively. On the other hand, some categories, such as

"Light" and "Bin," have much lower proportions, with only 0.24% and 0.18%, respectively.

Due to the reliance of neural networks on training data, class imbalance typically affects model performance [54]. In our analysis, we further demonstrated the impact of class proportions on segmentation performance, as illustrated in the Fig. 4, which shows the relationship between class pixel proportions and IoU scores for ConvNeXt and Swin Transformer when combined with the UperNet decoder. It can be observed that, in many cases, segmentation performance correlated positively with class proportion. Many classes, such as "Poster," "Light," and "Plant," were affected by their low pixel proportions, with lower segmentation performance observed for lower proportions. However, there are also classes that were less affected by pixel proportions, such as "Bin," which achieved an IoU score of 88.48% with ConvNeXt despite having only a 0.18% proportion. This is likely due to the distinct features of the "Bin" category, which makes it easier to segment.

Therefore, overcoming the influence of class imbalance is a critical challenge for improving model performance in indoor scene understanding.

B. Domain Gap

The pre-training and fine-tuning paradigm are a commonly used approach in semantic segmentation tasks, and from TABLE IV and TABLE V, it is evident that pre-training significantly enhances segmentation performance. However, the availability of pre-training datasets is currently limited, with no large-scale multispectral datasets available for pre-training. For segmentation tasks involving multispectral data, ImageNet is typically used as the pre-training dataset. However, ImageNet is an RGB dataset, and there is a significant channel domain gap between RGB and multispectral data [26]. For IndoorMS, which contains 9 channels, this channel difference poses a major challenge for model application.

This domain gap makes it challenging for the model to transfer learned certain features from RGB images to multispectral data due to mismatched channel characteristics [55]. Therefore, overcoming the channel domain gap between pre-training datasets and downstream task datasets is crucial for enhancing segmentation performance.

C. Limited Scale

The process of collecting multispectral data involves the use of multispectral sensors and requires data collection across numerous indoor environments. Moreover, data annotation requires manual semantic labeling of multichannel multispectral data, making the construction of the dataset costly. As a result, the scale of the IndoorMS dataset is limited, similar to other remote sensing image datasets, and is relatively small compared to traditional large-scale RGB datasets.

Nevertheless, the features contained in the IndoorMS dataset are extremely rich. Compared to RGB images, multispectral data can provide additional spectral channels, offering more detailed material and spectral characteristics, which can help the model capture information that is difficult to present in RGB data. This also presents an additional challenge for model learning, as it needs to effectively learn from these rich features despite the limited dataset size to improve segmentation performance in indoor scenes [56, 57].

VI. Conclusion and Outlook

This work presents IndoorMS, the first multispectral semantic segmentation dataset designed for indoor scene understanding, featuring fine-grained annotations for 19 semantic categories. We established benchmark performance for IndoorMS using a variety of representative semantic segmentation frameworks, with the best performance achieved by the combination of ConvNeXt and UperNet, yielding an $mIoU$ score of 72.9 and an $mF1$ score of 82.38. We analyzed the segmentation results of this dataset and identified key challenges, including class distribution imbalance, domain gap, and limited dataset scale. Based on these challenges, we propose the following directions for future research:

1. Training and Data Augmentation Strategies: Explore more effective training or data augmentation strategies to address class imbalance, particularly optimizing the performance for rare categories.
2. Pre-training Strategy for Domain Gap: Design a pre-training strategy based on ImageNet that can mitigate the channel domain gap between RGB and multispectral data during transfer learning.
3. Multispectral Pre-training Dataset: Construct a large-scale multispectral pre-training dataset to provide robust pre-trained models for tasks like multispectral semantic segmentation and object detection, addressing the limited options for pre-training datasets in multispectral analysis tasks.
4. Specialized Neural Networks for Multispectral Data: Design neural networks specifically for multispectral data to extract channel-specific features more effectively. Additionally, develop channel selection strategies to utilize the most effective spectral channels for training and inference.
5. Efficient Neural Networks for Few-shot Learning: Develop more efficient neural networks that can fully exploit the features in limited data settings, improving performance in small-sample training scenarios.
6. Indoor Multispectral Object Detection and Instance Segmentation: Future work can direct to collecting a multispectral dataset for indoor object detection [58] and instance segmentation, providing more comprehensive perception for indoor robots.
7. Integration of real and synthetic multispectral data: Future work could explore hybrid approaches that integrate real and synthetic multispectral data, combining the scalability of synthetic data with the fidelity of real data to enhance model generalization.

Acknowledgments

We would like to express our gratitude to Haoxing Yin, Long Qian, Chuhe Zhang, Yuhan Gao, Ningxin Weng, and Yuan Fang for their assistance in data collection and annotation.

REFERENCES

[1] W. Zhou, Y. Yue, M. Fang, X. Qian, R. Yang, and L. Yu, "BCINet: Bilateral cross-modal interaction network for indoor scene understanding in RGB-D images," *Information Fusion,* vol. 94, pp. 32-42, 2023, doi: 10.1016/j.inffus.2023.01.016.

[2] M. Roberts *et al.*, "Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10912-10922.

[3] C. Ye, Y. Yang, R. Mao, C. Fermüller, and Y. Aloimonos, "What can i do around here? deep functional scene understanding for cognitive robots," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*, 2017: IEEE, pp. 4604-4611, doi: 10.1109/ICRA.2017.7989535.

[4] B. Jia *et al.*, "Sceneverse: Scaling 3d vision-language learning for grounded scene understanding," in *European Conference on Computer Vision*, 2024: Springer, pp. 289-310.

[5] Z. B. Haladová, R. Szemző, T. Kovačovský, and J. Žižka, "Utilizing multispectral scanning and augmented reality for enhancement and visualization of the wooden sculpture restoration process," *Procedia Computer Science,* vol. 67, pp. 340-347, 2015, doi: 10.1016/j.procs.2015.09.278.

[6] R. Velastegui, M. Tatarchenko, S. Karaoglu, and T. Gevers, "Image semantic segmentation of indoor scenes: A survey," *Computer Vision and Image Understanding,* vol. 248, p. 104102, 2024, doi: 10.1016/j.cviu.2024.104102.

[7] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image segmentation using deep learning: A survey," *IEEE transactions on pattern analysis and machine intelligence,* vol. 44, no. 7, pp. 3523-3542, 2021, doi: 10.1109/TPAMI.2021.3059968.

[8] Y. Cai, L. Fan, and Y. Fang, "SBSS: Stacking-based semantic segmentation framework for very high-resolution remote sensing image," *IEEE Transactions on Geoscience and Remote Sensing,* vol. 61, pp. 1-14, 2023, doi: 10.1109/TGRS.2023.3234549.

[9] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431-3440.

[10] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, 2015: Springer, pp. 234-241, doi: 10.1007/978-3-319-24574-4_28.

[11] 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 transactions on pattern analysis and machine intelligence,* vol. 40, no. 4, pp. 834-848, 2017, doi: 10.1109/TPAMI.2017.2699184.

[12] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929,* 2020.

[13] Q. Zhu, Y. Fang, Y. Cai, C. Chen, and L. Fan, "Rethinking Scanning Strategies with Vision Mamba in Semantic Segmentation of Remote Sensing Imagery: An Experimental Study," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing,* 2024, doi: 10.1109/JSTARS.2024.3472296.

[14] Q. Zhu *et al.*, "Samba: Semantic segmentation of remotely sensed images with state space model," *Heliyon,* 2024, doi: doi.org/10.1016/j.heliyon.2024.e38495.

[15] X. Ma, X. Zhang, and M.-O. Pun, "RS3Mamba: Visual State Space Model for Remote Sensing Image Semantic Segmentation," *IEEE Geoscience and Remote Sensing Letters,* 2024, doi: 10.1109/LGRS.2024.3414293.

[16] S. I. Salem, H. Higa, J. Ishizaka, N. Pahlevan, and K. Oki, "Spectral band-shifting of multispectral remote-sensing reflectance products: Insights for matchup and cross-mission consistency assessments," *Remote Sensing of Environment,* vol. 299, p. 113846, 2023, doi: 10.1016/j.rse.2023.113846.

[17] Q. Zhu, Y. Cai, and L. Fan, "Seg-LSTM: Performance of xLSTM for Semantic Segmentation of Remotely Sensed Images," *arXiv preprint arXiv:2406.14086,* 2024.

[18] X. Ma, X. Zhang, X. Ding, M.-O. Pun, and S. Ma, "Decomposition-based Unsupervised Domain Adaptation for Remote Sensing Image Semantic Segmentation," *IEEE Transactions on Geoscience and Remote Sensing,* 2024, doi: 10.1109/TGRS.2024.3483283.

[19] M.-A. Ilișanu, F. Moldoveanu, and A. Moldoveanu, "Multispectral imaging for skin diseases assessment—state of the art and perspectives," *Sensors,* vol. 23, no. 8, p. 3888, 2023, doi: 10.3390/s23083888.

[20] Z. Chen *et al.*, "Using unmanned aerial vehicle multispectral data for monitoring the outcomes of ecological restoration in mining areas," *Land Degradation & Development,* vol. 35, no. 4, pp. 1599-1613, 2024, doi: 10.1002/ldr.5010.

[21] M. M. Amiri, D. W. Messinger, and T. R. Hanneken, "Colorimetric characterization of multispectral imaging systems for visualization of historical artifacts," *Journal of Cultural Heritage,* vol. 68, pp. 136-148, 2024, doi: 10.1016/j.culher.2024.05.014.

[22] H. M. Abdullah *et al.*, "Present and future scopes and challenges of plant pest and disease (P&D) monitoring: Remote sensing, image processing, and artificial intelligence perspectives," *Remote Sensing Applications: Society and Environment,* vol. 32, p. 100996, 2023, doi: 10.1016/j.rsase.2023.100996.

[23] S. G. Yel and E. Tunc Gormus, "Exploiting hyperspectral and multispectral images in the detection of tree species: A review," *Frontiers in Remote Sensing,* vol. 4, p. 1136289, 2023, doi: 10.3389/frsen.2023.1136289

[24] D. J. Wiersma and D. A. Landgrebe, "Analytical design of multispectral sensors," *IEEE Transactions on Geoscience and Remote Sensing,* no. 2, pp. 180-189, 1980, doi: 10.1109/TGRS.1980.350271.

[25] E. Tadmor, A. Nevet, G. Yahav, A. Fish, and D. Cohen, "Dynamic multispectral imaging using the vertical overflow drain structure," *IEEE Sensors Journal,* vol. 15, no. 7, pp. 3967-3972, 2015, doi: 10.1109/JSEN.2015.2406811.

[26] O. Glatt *et al.*, "Beyond RGB: a real world dataset for multispectral imaging in mobile devices," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 4344-4354.

[27] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, "Scene parsing through ade20k dataset," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 633-641.

[28] Q. Zhu, J. Cao, Y. Cai, and L. Fan, "Evaluating the Impact of Point Cloud Colorization on Semantic Segmentation Accuracy," *arXiv preprint arXiv:2410.06725,* 2024.

[29] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part V 12*, 2012: Springer, pp. 746-760.

[30] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, "Scannet: Richly-annotated 3d reconstructions of indoor scenes," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5828-5839.

[31] G. Paulin and M. Ivasic-Kos, "Review and analysis of synthetic dataset generation methods and techniques for application in computer vision," *Artificial intelligence review,* vol. 56, no. 9, pp. 9221-9265, 2023, doi: 10.1007/s10462-022-10358-3.

[32] W. Li *et al.*, "Interiornet: Mega-scale multi-sensor photo-realistic indoor scenes dataset," *arXiv preprint arXiv:1809.00716,* 2018.

[33] J. McCormac, A. Handa, S. Leutenegger, and A. J. Davison, "Scenenet rgb-d: Can 5m synthetic images beat generic imagenet pre-training on indoor segmentation?," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2678-2687.

[34] I. Goodfellow *et al.*, "Generative adversarial nets," *Advances in neural information processing systems,* vol. 27, 2014.

[35] F.-A. Croitoru, V. Hondru, R. T. Ionescu, and M. Shah, "Diffusion models in vision: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 2023, doi: 10.1109/TPAMI.2023.3261988.

[36] M. Alibani, N. Acito, and G. Corsini, "Multispectral satellite image generation using StyleGAN3," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing,* vol. 17, pp. 4379-4391, 2024, doi: 10.1109/JSTARS.2024.3356618.

[37] S. Sifnaios, I. Zorzos, G. Arvanitakis, F. K. Konstantinidis, G. Tsimiklis, and A. Amditis, "Exploration and Mitigation of the Impact of Lighting Conditions on Multi-spectral Image Classification," in *2023 IEEE International Conference on Imaging Systems and Techniques (IST)*, 2023: IEEE, pp. 1-6, doi: 10.1109/IST59124.2023.10355660.

[38] Z. Lin, Z. Tian, Q. Zhang, H. Zhuang, and J. Lan, "Enhanced Visual SLAM for Collision-Free Driving with Lightweight Autonomous Cars," *Sensors,* vol. 24, no. 19, p. 6258, 2024, doi: 10.3390/s24196258.

[39] Z. Lin, Q. Zhang, Z. Tian, P. Yu, and J. Lan, "DPL-SLAM: Enhancing Dynamic Point-Line SLAM through Dense Semantic Methods," *IEEE Sensors Journal,* 2024, doi: 10.1109/JSEN.2024.3373892.

[40] F. Liu, G. Li, S. Yang, W. Yan, G. He, and L. Lin, "Detection of heterogeneity on multi-spectral transmission image based on multiple types of pseudo-color maps," *Infrared Physics & Technology,* vol. 106, p. 103285, 2020, doi: 10.1016/j.infrared.2020.103285.

[41] A. Kirillov *et al.*, "Segment anything," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4015-4026.

[42] S. Hao, Y. Zhou, and Y. Guo, "A brief survey on semantic segmentation with deep learning," *Neurocomputing,* vol. 406, pp. 302-321, 2020, doi: 10.1016/j.neucom.2019.11.118.

[43] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770-778.

[44] Y. Zhang, S. Bader, and B. Oelmann, "A lightweight convolutional neural network model for concrete damage classification using acoustic emissions," in *2022 IEEE Sensors Applications Symposium (SAS)*, 2022: IEEE, pp. 1-6, doi: 10.1109/SAS54819.2022.9881386.

[45] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 11976-11986.

[46] Z. Liu *et al.*, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10012-10022.

[47] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2881-2890.

[48] T. Xiao, Y. Liu, B. Zhou, Y. Jiang, and J. Sun, "Unified perceptual parsing for scene understanding," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 418-434.

[49] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "Bisenet: Bilateral segmentation network for real-time semantic segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 325-341.

[50] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and efficient design for semantic segmentation with transformers," *Advances in neural information processing systems,* vol. 34, pp. 12077-12090, 2021.

[51] Q. Zhu, L. Fan, and N. Weng, "Advancements in point cloud data augmentation for deep learning: A survey," *Pattern Recognition,* p. 110532, 2024, doi: 10.1016/j.patcog.2024.110532.

[52] Q. Sellat, S. K. Bisoy, and R. Priyadarshini, "Semantic segmentation for self-driving cars using deep learning: a survey," in *Cognitive big data intelligence with a metaheuristic approach*: Elsevier, 2022, pp. 211-238.

[53] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems,* vol. 25, 2012.

[54] X. Guo, Y. Yin, C. Dong, G. Yang, and G. Zhou, "On the class imbalance problem," in *2008 Fourth international conference on natural computation*, 2008, vol. 4: IEEE, pp. 192-201, doi: 10.1109/ICNC.2008.871.

[55] Q. Zhu, N. Weng, L. Fan, and Y. Cai, "Enhancing Environmental Monitoring through Multispectral Imaging: The WasteMS Dataset for Semantic Segmentation of Lakeside Waste," *arXiv preprint arXiv:2407.17028,* 2024.

[56] C. Feng, Z. Cao, Y. Xiao, Z. Fang, and J. T. Zhou, "Multi-spectral template matching based object detection in a few-shot learning manner," *Information Sciences,* vol. 624, pp. 20-36, 2023, doi: 10.1016/j.ins.2022.12.067.

[57] Y. Cai, H. Huang, K. Wang, C. Zhang, L. Fan, and F. Guo, "Selecting optimal combination of data channels for semantic segmentation in city information modelling (CIM)," *Remote Sensing,* vol. 13, no. 7, p. 1367, 2021, doi: 10.3390/rs13071367.

[58] S. Ren, "Faster r-cnn: Towards real-time object detection with region proposal networks," *arXiv preprint arXiv:1506.01497,* 2015.

Qinfeng Zhu (Graduate Student Member, IEEE) received the degree of Master of Research in Computer Science from the University of Liverpool, Liverpool, U.K., in 2023, where he is currently working towards the Ph.D. degree in Computer Science.

His research interests mainly lie on deep learning, especially in multi-modal information fusion, 3D computer vision, semantic segmentation, and data augmentation.

Jingjing Xiao is a scholar who obtained her bachelor's and master's degrees from the School of Mechatronics Engineering and Automation at the National University of Defense Technology in China in 2010 and 2012, respectively. She further pursued her doctoral studies at the University of Birmingham in the United Kingdom. Currently, she serves as a director and senior engineer in the Bio-Med Informatics Research Centre & Clinical Research Centre, The Second Affiliated Hospital of the Army Medical University in China.

Lei Fan (Senior Member, IEEE) received the Ph.D. degree from the University of Southampton, Southampton, UK, in 2018. He is currently an Associate Professor within Department of Civil Engineering at Xi'an Jiaotong Liverpool University, Suzhou, China.

His main research interests include lidar and photogrammetry techniques, point cloud, machine learning, deformation monitoring, semantic segmentations, monitoring of civil engineering structures and geohazards.