2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

A Context-Enhanced Full-Resolution Floor Plan Segmentation Network for Topological Semantic Mapping

Zhengcai Cao*, *Senior Member, IEEE*, Yiyang Sun, Zhe Ma, and MengChu Zhou, *Fellow, IEEE*

***Abstract*— Topological semantic maps provide a practical solution to enhance indoor navigation for the Partially Sighted or Visually Impaired (PSVI). Segmenting indoor floor plans and extracting boundaries are key to constructing these maps. The existing methods exhibit low accuracy in segmentation. To achieve desired high segmentation accuracy, we introduce a Context-Enhanced Full-Resolution Network (CEFRN) for floor plan segmentation. It is designed to harness the shallow detailed features and inter-category contextual dependencies inherent in floor plans. CEFRN integrates modified residual blocks to capture the low-stage full-resolution features while maintaining its compactness. A position attention module is employed to refine the deep-stage contextual information. We also propose a two-dimensional deep supervision method to merge features from both stages, which significantly boosts the feature representation ability of CEFRN. Finally, a practical topological semantic mapping method for PSVI indoor navigation is introduced. Experimental results demonstrate that CEFRN's segmentation accuracy well exceeds the state-of-the-art methods'. It can be used to well support accurate topological semantic mapping.**

I. INTRODUCTION

As an abstract representation of an environment structure, topological semantic maps provide a novel global mapping solution to aid Partially Sighted or Visually Impaired (PSVI) individuals in navigating unfamiliar indoor scene. This approach reduces the reliance on dense prior maps [1], [2]. Floor plans are commonly available and contain information about unchanging structures, such as walls and doors [3]. The concept of using these floor plans to construct topological semantic maps thus arises. In this process, extracting layout structure information by segmenting floor plans with high accuracy has become crucial. While humans can easily interpret floor plans, their automatic processing poses significant challenges to computing devices.

Traditionally, manually interpreting rasterized floor plans requires significant labor, time, and financial resources. Early attempts at understanding these plans mainly relied on rule-based heuristics and basic image processing techniques [4]–[6]. However, the dependence on manually crafted features restricts their ability to generalize effectively.

Recent research has begun to explore deep learning methods for floor plan segmentation. Liu et al. [7] present a convolutional neural network to identify junctions in a floor plan and connect these junctions to locate walls. Yet it falters with non-Manhattan layouts. Zeng et al. [8] devise a multi-task framework utilizing a room-boundary guided attention mechanism for predicting room types and boundary elements. However, this strategy improves room type predictions without significantly benefiting boundary element segmentation. Subsequent efforts in this field [9]–[12] capitalize on more sophisticated networks, but advancements in predicting generalizable room boundaries are minimal.

General image-based semantic segmentation networks like DeepLabV3+ [13], OCRNet [14] and HRNet [15] perform well in standard application scenarios [16]–[18] and can also be applied to floor plan segmentation. However, their limitation in capturing the unique detailed features and context-dependent information of floor plans hinders their performance. On the other hand, derived from the foundational U-Net [19] known for its encoder-decoder architecture and skip-connections, U-shaped segmentation networks [20]–[22] are able to preserve high-resolution low-level texture features. However, they lack the effective representation of the inter-category contextual dependencies in a floor plan.

After examining the existing work, we have identified that low-contrast areas in floor plans, such as doors, are the primary factors limiting the accuracy of floor plan segmentation when existing methods are directly deployed. This limitation arises from the lack of prominent, distinguishable features within these areas. Thus, it becomes essential to exploit the inter-category contextual relationships during segmentation. This specifically involves leveraging prior knowledge that doors are typically embedded within walls. Harnessing this insight opens a new way to improve segmentation accuracy. Moreover, the issue of class imbalance within floor plans also emphasizes the importance of detailed texture features. High-resolution feature processing emerges as a direct and effective strategy for reducing detail loss due to its ability to maintain finer image textures. However, existing methods [13]–[15], [20]–[22] for floor plan segmentation fail to simultaneously exploit the inherent inter-category contextual dependencies and high-resolution detailed features, thus limiting the segmentation accuracy.

To address this issue, we introduce a Context-Enhanced Full-Resolution Network (CEFRN), which is expanded with modified residual blocks. Its compact structure ensures the

This work is supported in part by the National Natural Science Foundation of China under Grant (92148202, 52175002), and the Beijing Natural Science Foundation (L223019, 3242011). (Corresponding authors: Zhengcai Cao.)

Zhengcai Cao is with State Key Laboratory of Robotics and Systems, Harbin Institute of Technology, Harbin, 150080, China (e-mail: caozc@hit.edu.cn)

Yiyang Sun and Zhe Ma are with the College of Information Science and Technology, Beijing University of Chemical Technology, Beijing, 100029, China (e-mail: peaksyy@163.com and mz0810@163.com)

MengChu Zhou is with the Department of Electrical and Computer Engineering, New Jersey Institute of Technology, Newark, NJ 07102 USA (e-mail: zhou@njit.edu).

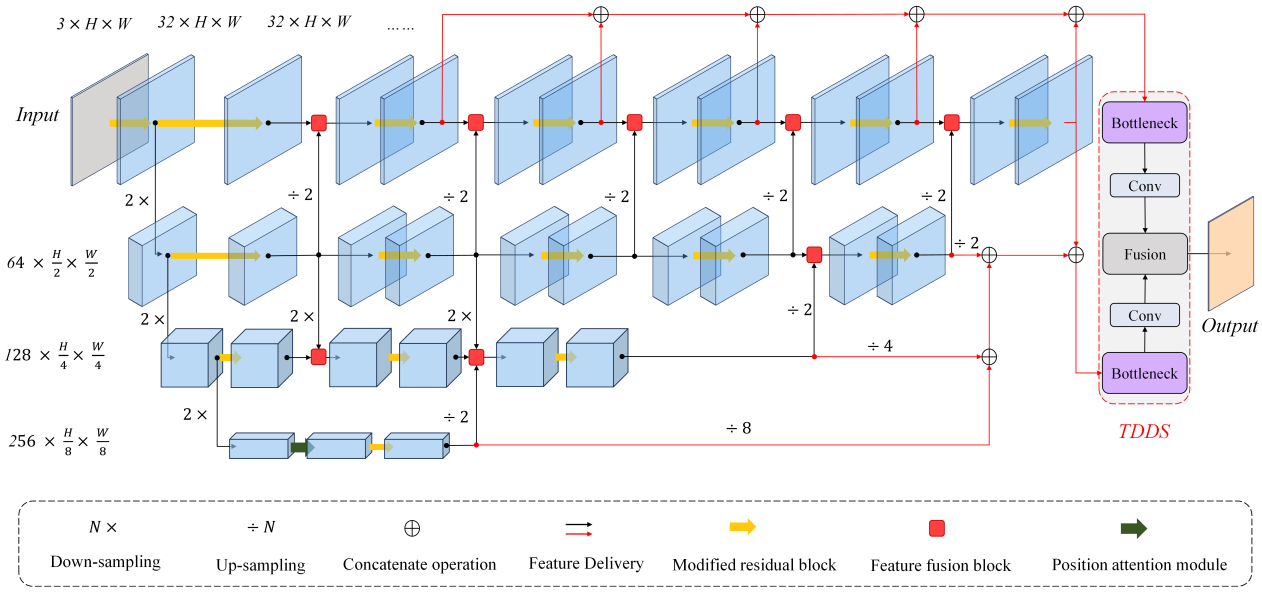


Fig. 1. The overall architecture of our context-enhanced full-resolution network. The area included by red lines represents TDDS. Details of the modified residual block and the feature fusion block are depicted as shown in Fig. 2 and Fig. 3, respectively.

capture of full-resolution texture details in floor plans. A position attention module is incorporated at the deepest stage of CEFRN to enhance the inter-category contextual information. We propose a Two-Dimensional Deep Supervision (TDDS) method to fuse the two types of crucial features at various scales, thus improving the network's representation capabilities. Moreover, we develop a topological semantic mapping technique. It utilizes the segmentation result to create a map that assists PSVI individuals in navigating unfamiliar indoor environments.

This work intends to make the following new contributions to the field of floor plan segmentation:

1) We develop a U-shaped network, CEFRN, for floor plan segmentation for the first time. It includes the modified residual blocks to expand the receptive field across multiple resolutions, while ensuring the network's compactness;
2) We propose TDDS that effectively integrates full-resolution detailed features and multi-scale contextual information. Its use is able to enhance CEFRN's feature representation capability; and
3) We propose a topological semantic mapping method by using floor plan segmentation mask as input, which eliminates the need for accurate prior maps in existing PSVI assistance navigation strategies.

The next section presents the details of our proposed method. Section III gives the experimental settings and results to compare the proposed method and the state of the art. Section IV concludes this paper.

II. PROPOSED METHOD

In this section, we first introduce the overall architecture of our compact yet effective CEFRN, designed for the semantic segmentation of two-dimensional rasterized floor plans. Then, we propose a topological semantic mapping method that leverages segmentation results to construct topological semantic maps, which can assist PSVI individuals in indoor navigation.

A. Context-Enhanced Full-Resolution Network (CEFRN)

1) Overall network architecture: As illustrated in Fig. 1, the proposed network architecture is built on the U-shaped encoder-decoder framework. We expand each stage of the network horizontally and vertically by up-sampling, down-sampling, and convolution. Cross-stage connections are added within both shallow and deep layers. This allows our network to enhance high-resolution features in shallow layers and contextual information in deep layers. The former can provide finer edge texture information, while the latter can complement inter-category dependencies. Given the architectural features mentioned above, this network is well-suited for full-resolution pixel-level prediction under high-level contextual conditions. In a floor plan segmentation task, it can also mitigate the loss of small target pixels during the down-sampling process.

A modified residual block is introduced for bidirectional expansion of CEFRN, as shown in Fig. 2. It utilizes two and three 3×3 convolution blocks to effectively approximate 5×5 and 7×7 convolution operations respectively. Dropout layer with a dropout rate of 0.2 is added after each Batch Normalization layer to reduce overfitting. This allows for multi-resolution feature analysis with reduced memory requirements, thus facilitating the model's practical deployment.

In CEFRN, the feature fusion block integrates the feature map from the preceding residual block with both up-sampled and down-sampled features from neighboring stages, illus-

trated in Fig. 3. The operations are defined as follows:

$$C_{r,c} = \begin{cases} D(x_{r-1,c}) \oplus U(x_{r+1,c}) \oplus (x_{r,c}), & \text{case 1} \\ D(x_{r-1,c}) \oplus (x_{r,c}), & \text{case 2} \\ U(x_{r+1,c}) \oplus (x_{r,c}), & \text{case 3} \end{cases} \quad (1)$$

where $D(\cdot)$ and $U(\cdot)$ represent down-sampling and up-sampling operation, respectively. $x_{r,c}$ denotes a stacked feature map at row r and column c in CEFRN. The operation is divided into three cases according to the location of the node in the network. $x_{r,c}$ is concatenated with the up-sampled output of $x_{r+1,c}$ and down-sampled output of $x_{r-1,c}$. The resulting $C_{r,c}$ is then passed through a parallel sequence including a 1×1 convolution, 3×3 convolution and 3×3 atrous convolution with a dilation rate of 2.

Deep network layers are rich in high-level semantic information. The position attention module [23] has demonstrated its ability to enhance the representation capability of contextual information within local features. We thus embed it in the deepest stage of CEFRN, as shown in Fig. 1. This enables the position attention module to effectively utilize high-level semantic information and learn the inter-category positional dependencies.

2) TDDS: For floor plans, shallow full-resolution features and deep aggregated features are crucial since they can offer low-level details and high-level context, respectively. To harness both aspects, we propose the TDDS across the network, as shown in Fig. 1. This process involves fusing each feature layer $\{f_1, ..., f_n\}$ from the first stage of the network, where n denotes the number of semantic levels [20] in this stage. This is followed by processing through a bottleneck, which include a 3×3 convolution, batch normalization, and ReLU activation function. Subsequently, a 1×1 convolution layer is employed to generate a probability map. Each decoder stage then produces side output $\{s_1, ..., s_i\}$ via up-sampling, where i signifies the number of side outputs. The same deep supervision mechanism is applied to create the corresponding probability maps. The generated probability maps are defined as:

$$P_f = \mathbf{Conv_{1\times1}}(\mathbf{B_N}(f_1 \oplus \ldots \oplus f_n)), \quad (2)$$

$$P_s = \mathbf{Conv_{1\times1}}(\mathbf{B_N}(s_1 \oplus \ldots \oplus s_i)), \quad (3)$$

where $\mathbf{B_N}(\cdot)$ denotes a bottleneck layer, $\mathbf{Conv_{1\times1}}(\cdot)$ refers to a 1×1 convolution operation. P_f and P_s are the probability maps from low-level detailed dimension and high-level context dimension, respectively.

The final predicted probability output is derived as the weighted sum of these probability maps, which is calculated as follows:

$$P = \omega_1 \cdot P_f + \omega_2 \cdot P_s, \quad (4)$$

where ω_1 and ω_2 represent output weights in different deep supervision dimension.

Moreover, recall loss [24] is introduced as a loss function to mitigate the class imbalance issue. In floor plan images, the number of background pixels typically surpasses that of wall pixels, which in turn exceed the count of door pixels. Given the disparity in pixel counts across different classes, balancing their contributions in the segmentation task becomes crucial. Compared with weighted cross entropy loss and focal loss [25], recall loss can also play a role in enhancing positive sample identification and reducing missed detections. It weights the standard cross entropy loss for each class with its instantaneous training recall performance. It is defined as follows:

$$Loss = -\sum_{c=1}^{C} \sum_{n:y_i=c} \left(1 - \frac{T_P^{c,t}}{F_N^{c,t} + T_P^{c,t}}\right) \log(p_{n,t}), \quad (5)$$

where $T_P^{c,t}$ and $F_N^{c,t}$ indicate the true positive and false negative count for class c at optimization step t, respectively. $n : y_i = c$ represents all samples such that the ground truth label y_i is class c. $p_{n,t}$ is the predicted probability for sample n at time t.

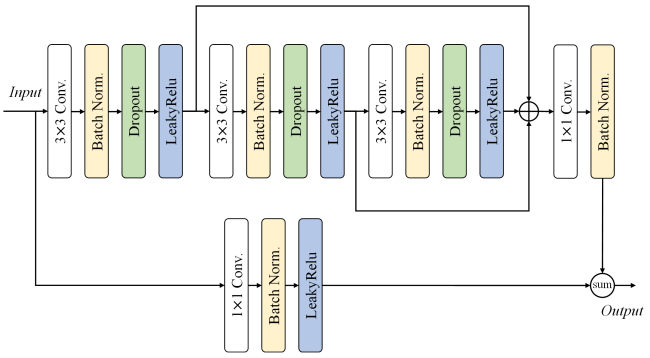


Fig. 2. Illustration of a modified residual block. Various cascaded combinations of 3×3 convolutions are embedded into the residual module to expand the receptive field of the convolutional module with fewer parameters.

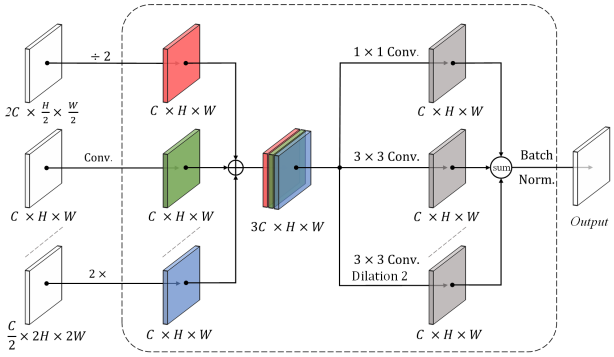
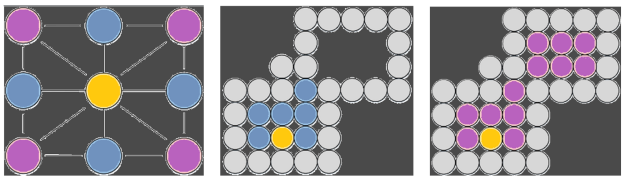


Fig. 3. Illustration of a feature fusion block. The up-sampling and down-sampling feature maps from adjacent stages are integrated by using concatenation and parallel multi-convolutions.

B. Topological Semantic Mapping

In this part, we present a method to construct a topological semantic map from a floor plan. This method comprises two parts: topology calculation and semantic matching.

1) Topology calculation: As an abstract representation, a topological semantic map consists of nodes and edges representing places and their navigable connections, respectively. Nodes in floor plans include rooms, doors, and intersections



(a) (b) (c)

Fig. 4. Schematic diagrams of the seed fill algorithm. The white, blue, and purple areas represent walls, four-connectivity adjacent pixels, and eight-connectivity adjacent pixels, respectively.

in the corridors. Building such a map requires identifying isolated places as nodes. Initially, floor plan segmentation result is used to extract semantic masks via threshold segmentation. Dilation and erosion are then applied to remove noise, which ensures accuracy in subsequent steps. The denoised masks M_r and M_d are input into in Algorithm 1 to calculate room nodes. After that, the center coordinates and pixel values of each room are stored in a dictionary D_r, and re-labeled room areas in mask M_{lr}.

Algorithm 1 Room node calculation

Input: M_r, M_d: Denoised room and door masks with $W \times H$

Output: D_r: Room dictionary, M_{lr}: Re-labeled room mask

δ: Random value from 1 to 254
$SeedFill(\alpha, \beta)$: All pixels in mask α that are adjacent to pixel β in four directions and have the same value
$Contour(\alpha)$: Extract the outer contour of area α
$Center(\alpha)$: Calculate the center coordinates of contour α
$L_r \leftarrow 1 \quad M \leftarrow M_r + M_d$
for $i = 1$ to W **do**
  for $j = 1$ to H **do**
    if $M(i, j)$ is 0 **then**
      $M_{lr}(i, j) \leftarrow \delta$
      $M_{lr}(SeedFill(M, (i, j))) \leftarrow \delta$
    end if
    $R \leftarrow$ All pixels in M with value δ
    $p \leftarrow Center(Contour(R))$
    $D_r(L_r) \leftarrow (\delta, p)$
    Choose a new value for δ within range 1 to 254
    Increment L_r by 1
  end for
end for

Algorithm 1 employs a seed fill algorithm to label adjacent pixels with the same value, thus identify different rooms. It is divided into four-connectivity and eight-connectivity types, based on neighbor selection methods, depicted in Fig. 4. Fig. 4(a) shows the distribution of adjacent pixels for both types. We opt for the four-connectivity approach to avoid scenarios like those shown in Fig. 4(c) and ensure distinct labels for each room.

Another crucial element in the floor plan is doors. Similar to Algorithm 1, the denoised door mask M_d is used to extract and label the coordinates of doors separately, which are then stored as D_d. For each door in the floor plan, we retrieve the adjacent pixel values in M_{lr} using the door boundary pixels as seed. We query the dictionary D_r to obtain information on adjacent rooms, and then construct the adjacency matrix $\mathbb{M}$. In addition to room and door nodes, intersections in corridor are identified using the method detailed in [26]. Empirically, a room area without text is considered as the corridor. After abstracting the three types of nodes, we connect the center of each room to the centers of its corresponding doors as indoor edges, based on the matrix $\mathbb{M}$. Furthermore, each door node is connected to the nearest one or two intersection nodes in the corridor, which serves as branch edges. Each intersection node in the corridor is linked to the nearest one or two intersection nodes in different directions, which serves as trunk edges.

2) Semantic matching: In addition to nodes and edges, another important element of a topological semantic map is the semantics of nodes. Text is often used in floor plans to identify area types, thus becoming a source of semantic information. In our work, PP-OCRV3 [27] is employed to detect and recognize text in floor plans. The output of the model is a list of text center coordinates and their content. We retrieve the pixel value of the re-labeled room mask at the center coordinate of each text. This step is performed to obtain the corresponding room label from the room dictionary D_r. Each text is then matched with its corresponding room, which provides semantic information to the room. In practical use, we substitute auto-calculated room centers with text coordinates under specific circumstances.

III. EXPERIMENTAL RESULTS AND ANALYSIS

A. Dataset and Evaluation Metrics

1) Dataset: Recent corporate restrictions have resulted in the unavailability of raw images from several floor plan datasets. Given the high cost of manual annotation, most datasets lack comprehensive annotations. Among these, the Rent3D (R3D) [30] is notable for being the most comprehensive floor plan dataset publicly available, featuring pixel-level annotations for semantic segmentation. Specifically, R3D contains 214 original images, most of which showcase irregular room shapes and nonuniform wall thicknesses. Following [8], an additional 18 images of round-shaped floor plans are incorporated into R3D. All experiments in this section are conducted on R3D. For a fair comparison, we adhere to the split proposed in the original paper [8], allocating 179 images for training and 53 for testing.

2) Evaluation Metrics: Following [8], [10], [12], we employ Intersection over Union (IoU), Precision (Pre), Recall (Rec), and F-measure [31] for quantitative evaluation:

$$IoU = \frac{T_P}{T_P + F_N + F_P}, \tag{6}$$

$$Pre = \frac{T_P}{T_P + F_P}, \tag{7}$$

TABLE I

COMPARISON RESULT ON THE TEST SET. THE BEST SCORES ARE HIGHLIGHTED IN BOLD, WHILE THE SECOND BEST ARE UNDERLINED.

Method	Backbone	Param (M)	Background				Wall				Door			
			IoU	Pre	Rec	F_β	IoU	Pre	Rec	F_β	IoU	Pre	Rec	F_β
DFPR [8]	VGG-16	28.91	0.9822	**0.9975**	0.9847	0.9945	0.8089	0.8199	**0.9837**	0.8527	0.6068	0.6996	0.8207	0.7243
UNet++ [20]	-	8.37	0.9850	0.9913	0.9940	0.9919	0.8419	0.9185	0.9099	0.9165	0.6082	0.8215	0.7008	0.7902
UNet3+ [21]	-	26.99	0.9865	0.9942	0.9923	0.9938	0.8542	0.9001	0.9437	0.9098	0.6312	0.8136	0.7379	0.7948
DeepLabV3+ [13]	ResNet-101	45.83	0.9866	0.9939	0.9923	0.9935	0.8366	0.9004	0.9222	0.9053	0.6703	0.8038	0.8014	0.8032
GCNet [28]	ResNet-101	68.72	0.9812	0.9941	0.9869	0.9924	0.7804	0.8443	0.9115	0.8589	0.6747	0.7598	0.8576	0.7803
OCRNet [14]	HRNet-W48	70.45	0.9753	0.9922	0.9829	0.9900	0.7202	0.7817	0.9016	0.8066	0.6815	0.8400	0.7831	0.8261
HRNet-Contrast [15]	HRNet-W48	70.08	0.9856	0.9905	0.9950	0.9915	0.8338	0.9252	0.8941	0.9178	0.6380	0.8478	0.7206	0.8146
SegNeXt-L [29]	MSCAN-L	45.11	0.9849	0.9923	0.9925	0.9923	0.8245	0.9067	0.9009	0.9054	0.6541	0.7864	0.7955	0.7885
CEFRN (Ours)	-	**4.40**	**0.9872**	0.9948	**0.9952**	**0.9949**	**0.8546**	**0.9340**	0.9689	**0.9418**	**0.7102**	**0.8823**	**0.8696**	**0.8793**

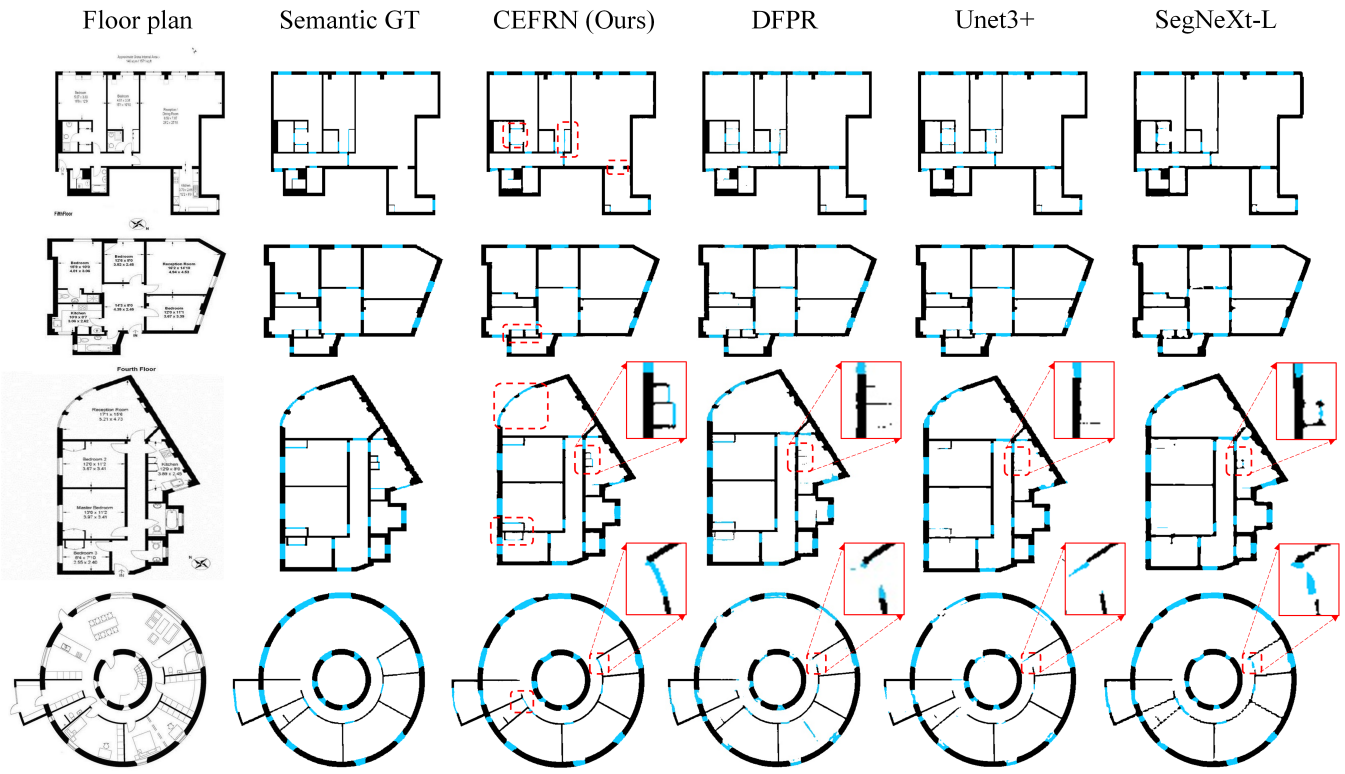


Fig. 5. Visual comparison of floor plan segmentation results. Black pixels represent walls, blue pixels represent doors, and the red dashed box highlights our CEFRN's visualization enhancement.

$$Rec = \frac{T_P}{T_P + F_N}, \tag{8}$$

where True Positive (T_P) represents correctly predicted positive pixels, False Positive (F_P) indicates pixels incorrectly predicted as positive, and False Negatives (F_N) denote missed positive pixels.

F-measure incorporates both Pre and Rec, providing a unified standard for imbalanced datasets. This facilitates a more direct and fair evaluation of performance. It is expressed as:

$$F_\beta = \frac{(1+\beta^2)\text{Pre} \times \text{Rec}}{\beta^2 \text{Pre} + \text{Rec}}, \tag{9}$$

where β is the configuration parameter that reconciles Pre and Rec.

B. Implementation Details

All the experiments are carried out on a desktop with NVIDIA RTX3090 GPU on Ubuntu system. The Adam optimizer is used for parameter updates with a fixed learning rate of 0.0001 to train the model. We limit the batch size to 1 without employing any training tricks and run 40k iterations in total. Input images are resized to 512×512 to preserve the thin and short lines in the floor plans. We train comparison methods using hyper-parameters from their original papers. Empirically, we set β^2 to 0.3 and the ω_1 and ω_2 in TDDS each to 0.5.

C. Comparison with Segmentation Networks

We first conduct a comparison between our method and the widely recognized floor plan segmentation method DFPR [8]. For a fair evaluation, we adopt the same training setting in the DFPR's original paper. We next compare our method with high-performing U-shaped image segmentation networks including UNet++ [20] and UNet3+ [21]. Finally, we assess our network's performance with existing state-of-the-art (SOTA) image-based semantic segmentation methods, e.g., DeepLabV3+ [13], OCRNet [14], and SegNeXt-L [29]. The backbones used in these methods, including VGG-16, ResNet101, HRNet-W48, and MSCAN-L are pretrained on ImageNet [32].

The experimental results are reported in Table I. CEFRN ranks as the best or second-best across all metrics, surpassing SOTA models in overall performance. Notably, in segmenting the challenging door category, it significantly outperforms the second-best OCRNet [14], with a 4.21% increase in IoU and a 6.43% increase in F_β. The effective cross-stage feature integration capability of TDDS is the main reason for the improvement in segmentation accuracy. Furthermore, our model has achieved the minimal parameter count of 4.4 million, bolstering its deployability. Compared to OCRNet, which ranks second in door category segmentation performance, our model's parameter count is only 6.25% of theirs. This is attributed to the modified residual block in CEFRN, which uses cascaded 3×3 convolutions to approximate larger 5×5 and 7×7 convolutions, thus expanding the receptive field while maintaining a lower parameter count.

From Table I, it can be seen that Pre of the background class and Rec of the wall class from DFPR [8] exceed those of ours. Higher Pre implies fewer false positive pixels, and higher Rec indicates fewer false negative pixels, according to (7) and (8). Despite this, our method performs better than DFPR on the F_β, which prioritizes overall performance.

The visualization in Fig. 5 illustrates CEFRN's superior performance over its representative peers [8], [21], [29]. Row 3 shows that CEFRN can still achieve outstanding results in segmenting small targets compared to its peers. This can be attributed to its full-resolution and multi-scale representation capabilities, which makes it more sensitive to detailed textures. Row 4 demonstrates that our method surpasses others in door segmentation within areas where the wall is almost entirely segmented. This is due to CEFRN integrates deep contextual features, enabling it to deduce door positions from walls based on inter-category dependencies.

TABLE II

ABLATION STUDY ON CEFRN. TDDS, MRB, PAM REPRESENT TWO-DIMENSIONAL DEEP SUPERVISION, MODIFIED RESIDUAL BLOCK, AND POSITION ATTENTION MODULE, RESPECTIVELY.

Model	TDDS	MRB	PAM	Param (M)	$mIoU$	mF_β
Model-0	×	×	×	14.88	0.8010	0.8774
Model-T	✓	×	×	15.14	0.8365	0.9133
Model-C	×	✓	×	4.00	0.8183	0.8934
Model-M	×	×	✓	15.03	0.8250	0.9099
Model-TC	✓	✓	×	4.25	0.8319	0.9149
Model-CM	×	✓	✓	4.14	0.8201	0.9012
Model-TM	✓	×	✓	15.28	**0.8515**	0.9379
CEFRN	✓	✓	✓	4.40	0.8507	**0.9387**

TABLE III

COMPARISON OF THE IMPACT OF DIFFERENT SEGMENTATION MODELS ON TOPOLOGICAL SEMANTIC MAPPING PROCESS. † REPRESENTS THE TOPOLOGICAL SEMANTIC MAPPING BASED ON THIS SEGMENTATION MODEL.

Method	Rooms		Doors		Intersections	
	$C_{T_P} \uparrow$	$C_{F_P} \downarrow$	$C_{T_P} \uparrow$	$C_{F_P} \downarrow$	$C_{T_P} \uparrow$	$C_{F_P} \downarrow$
DFPR†	219/221	6/221	265/272	25/272	74/85	5/85
UNet++†	206/221	7/221	249/272	11/272	82/85	3/85
UNet3+†	208/221	5/221	257/272	13/272	79/85	5/85
DeepLabV3+†	210/221	7/221	263/272	16/272	79/85	5/85
GCNet†	203/221	8/221	268/272	20/272	75/85	7/85
OCRNet†	200/221	10/221	255/272	6/272	69/85	9/85
HRNet-Contrast†	218/221	**0/221**	257/272	4/272	81/85	4/85
SegNeXt-L†	217/221	1/221	263/272	17/272	78/85	4/85
CEFRN† (Ours)	**221/221**	**0/221**	**270/272**	**2/272**	**83/85**	**2/85**

D. Ablation Study

To further reveal the contribution of each component in our method, we conduct an ablation study on CEFRN. We employ average category-wise Intersection over Union (mIoU) and average category-wise F-measure (mF_β) as quantitative metrics, and give the results in Table II.

To ensure comparability, parallel 3×3, 5×5 and 7×7 convolution blocks with a residual connection are used to replace the modified residual block in Model-TM. Table II shows that CEFRN, though slightly behind Model-TM in performance, benefits from the modified residual block, which reduces parameters by 10.88 million from TM's. This reduction lowers memory usage, maintains the compactness of the network, and improves the feasibility of real-world deployment. In

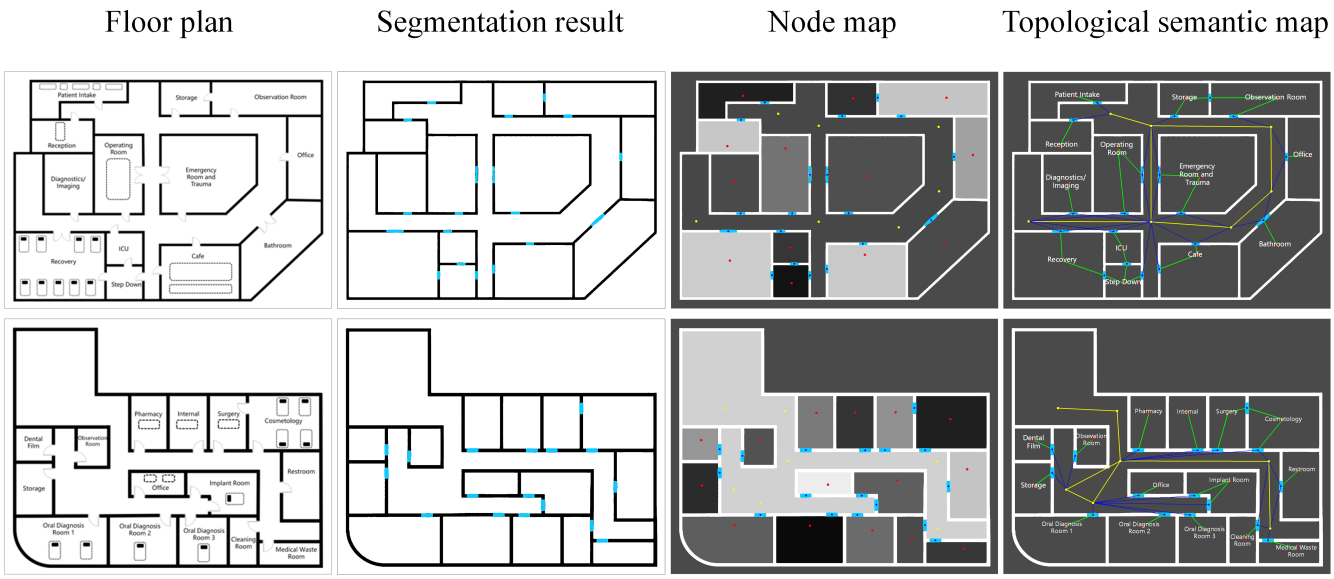


Fig. 6. Experimental results of topological semantic mapping. The first column shows original hospital floor plans. The second column details the segmentation results output by CEFRN. The third column shows the topological node maps with red, blue, and yellow points representing rooms, doors, and intersections, respectively. The fourth column shows the topological semantic maps, where green, blue, and yellow lines represent indoor, branch, and trunk edges, respectively.

Model-CM, the TDDS component is removed, the output is redirected to the final feature layer of the first stage after a 1×1 convolution operation. Compared to Model-CM, CEFRN exhibits a 3.73% increase in mIoU and a 4.16% improvement in mF_β. This is attributed to TDDS, which deeply integrates low-stage details and multi-scale context features, enhancing CEFRN's ability to represent floor plans. In Model-TC, removing the positional attention module from the last stage shows that including this module brings a 2.25% improvement in mIoU and a 2.60% improvement in mF_β for CEFRN. This is due to the module's enhanced ability to represent contextual information in local features, which strengthen the inter-category semantic dependencies. By analogy, Model-T, Model-C and Model-M demonstrate their contributions in comparison with the baseline Model-0.

E. CEFRN Evaluation for Topological Semantic Mapping

To construct a topological semantic map, the segmentation model is essential for extracting the boundary elements of a floor plan. We compare the effects of different segmentation models on topological semantic mapping to further reveal the superiority of CEFRN. To the best of our knowledge, the labeled floor plans in existing datasets are residential layouts. However, the real scenes PSVI individuals encounter are public places, such as hospitals. We have thus collected 17 hospital floor plans for evaluation.

We use the number of true positives C_{T_P} and false positives C_{F_P} in calculating room, door, and intersection nodes to quantify the impact of segmentation models on topological semantic mapping. As reported in Table III, CEFRN's performance significantly surpasses its peers', highlighting its exceptional enhancement capabilities in topological semantic mapping and robust generalization in hospital settings. Besides the segmentation results, topological semantic mapping is also influenced by factors such as the clarity of floor plans, which is not further detailed in this paper. The CEFRN-based topological semantic mapping processes for two hospital floor plans are shown in Fig. 6.

IV. CONCLUSIONS

In this work, we introduce an accurate and compact context-enhanced full-resolution network for floor plan segmentation. This network addresses the issue of low segmentation accuracy, which is attributed to existing methods' underestimation of the shallow texture features and concealed inter-category dependencies within the floor plans. Specifically, it expands parallel convolutional layers both horizontally and vertically by using modified residual blocks. Additionally, a position attention module is used to enhance inter-category dependencies, while a two-dimensional deep supervision method is proposed to integrate horizontal high-resolution texture and vertical multi-scale context in CEFRN. A topological semantic mapping method is introduced to construct topological semantic maps for indoor navigation based on CEFRN. Quantitative evaluations show that CEFRN surpasses the state of the art and can be used to obtain highly accurate topological semantic mapping. Our next work plans to focus on: 1) creating a labeled hospital floor plan dataset to refine and adapt CEFRN for common scenarios encountered by PSVI people, and 2) applying our novel method to mobile robots, autonomous vehicles, and drone navigation [33]–[39].

REFERENCES

[1] X. Chen, B. Zhou, J. Lin, Y. Zhang, F. Zhang, and S. Shen, "Fast 3d sparse topological skeleton graph generation for mobile robot global planning," in *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, pp. 10283–10289, 2022.

[2] H. Liu, H. Huang, S.-K. Yeung, and M. Liu, "360st-mapping: An online semantics-guided topological mapping module for omnidirectional visual slam," in *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, pp. 802–807, 2022.

[3] N. Zimmerman, M. Sodano, E. Marks, J. Behley, and C. Stachniss, "Constructing metric-semantic maps using floor plan priors for long-term indoor localization," in *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, pp. 1366–1372, 2023.

[4] P. Dosch, K. Tombre, C. Ah-Soon, and G. Masini, "A complete system for the analysis of architectural drawings," *Int. Journal on Document Analysis and Recognition*, vol. 3, no. 2, pp. 102–116, 2000.

[5] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (surf)," *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 346–359, 2008.

[6] S. Ahmed, M. Liwicki, M. Weber, and A. Dengel, "Automatic room detection and room labeling from architectural floor plans," in *Int. Workshop on Document Analysis Systems*, pp. 339–343, IEEE, 2012.

[7] C. Liu, J. Wu, P. Kohli, and Y. Furukawa, "Raster-to-vector: Revisiting floorplan transformation," in *Proc. of the IEEE Int. Conf. on Computer Vision (ICCV)*, pp. 2195–2203, 2017.

[8] Z. Zeng, X. Li, Y. K. Yu, and C.-W. Fu, "Deep floor plan recognition using a multi-task network with room-boundary-guided attention," in *Proc. of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9096–9104, 2019.

[9] I. Y. Surikov, M. A. Nakhatovich, S. Y. Belyaev, and D. A. Savchuk, "Floor plan recognition and vectorization using combination unet, faster-rcnn, statistical component analysis and ramer-douglas-peucker," in *Int. Conf. on Computing Science, Communication and Security*, pp. 16–28, Springer, 2020.

[10] X. Lv, S. Zhao, X. Yu, and B. Zhao, "Residential floor plan recognition and reconstruction," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 16717–16726, 2021.

[11] Z. Lu, T. Wang, J. Guo, W. Meng, J. Xiao, W. Zhang, and X. Zhang, "Data-driven floor plan understanding in rural residential buildings via deep recognition," *Information Sciences*, vol. 567, pp. 58–74, 2021.

[12] B. Yang, H. Jiang, H. Pan, and J. Xiao, "Vectorfloorseg: Two-stream graph attention network for vectorized roughcast floorplan segmentation," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 1358–1367, 2023.

[13] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. of the European Conf. on Computer Vision (ECCV)*, pp. 801–818, 2018.

[14] Y. Yuan, X. Chen, and J. Wang, "Object-contextual representations for semantic segmentation," in *Proc. of the European Conf. on Computer Vision (ECCV), pages=173–190, year=2020*.

[15] W. Wang, T. Zhou, F. Yu, J. Dai, E. Konukoglu, and L. Van Gool, "Exploring cross-image pixel contrast for semantic segmentation," in *Proc. of the IEEE/CVF Int. Conf. on Computer Vision (ICCV)*, pp. 7303–7313, 2021.

[16] S. Kuutti, R. Bowden, Y. Jin, P. Barber, and S. Fallah, "A survey of deep learning applications to autonomous vehicle control," *IEEE Trans. on Intelligent Transportation Systems*, vol. 22, no. 2, pp. 712–733, 2020.

[17] G. Roggiolani *et al.*, "On domain-specific pre-training for effective semantic perception in agricultural robotics," in *IEEE Int. Conf. on Robotics and Automation (ICRA)*, pp. 11786–11793, IEEE, 2023.

[18] C. Wang, W. Pedrycz, Z. Li, and M. Zhou, "Residual-driven fuzzy c-means clustering for image segmentation," *IEEE/CAA Journal of Automatica Sinica*, vol. 8, no. 4, pp. 876–889, 2020.

[19] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Int. Conf., Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pp. 234–241, Springer, 2015.

[20] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: Redesigning skip connections to exploit multiscale features in image segmentation," *IEEE Trans. on Medical Imaging*, vol. 39, no. 6, pp. 1856–1867, 2019.

[21] H. Huang, L. Lin, R. Tong, H. Hu, Q. Zhang, Y. Iwamoto, X. Han, Y.-W. Chen, and J. Wu, "Unet 3+: A full-scale connected unet for medical image segmentation," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Process (ICASSP)*, pp. 1055–1059, IEEE, 2020.

[22] N. Ibtehaz and M. S. Rahman, "Multiresunet: Rethinking the u-net architecture for multimodal biomedical image segmentation," *Neural networks*, vol. 121, pp. 74–87, 2020.

[23] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 3146–3154, 2019.

[24] J. Tian, N. C. Mithun, Z. Seymour, H.-p. Chiu, and Z. Kira, "Recall loss for imbalanced image classification and semantic segmentation," 2021.

[25] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. of the IEEE Int. Conf. on Computer Vision (ICCV)*, pp. 2980–2988, 2017.

[26] Y. Sun, Z. Ma, M. Zhou, and Z. Cao, "A topological semantic mapping method based on text-based unsupervised image segmentation for assistive indoor navigation," *IEEE Trans. on Instrumentation and Measurement*, vol. 72, pp. 1–13, 2023.

[27] C. Li *et al.*, "Pp-ocrv3: More attempts for the improvement of ultra lightweight ocr system," *arXiv preprint arXiv:2206.03001*, 2022.

[28] Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu, "Gcnet: Non-local networks meet squeeze-excitation networks and beyond," in *Proc. of the IEEE/CVF Int. Conf. on Computer Vision Workshops*, 2019.

[29] M.-H. Guo, C.-Z. Lu, Q. Hou, Z. Liu, M.-M. Cheng, and S.-M. Hu, "Segnext: Rethinking convolutional attention design for semantic segmentation," *Advances in Neural Information Processing Systems*, vol. 35, pp. 1140–1156, 2022.

[30] C. Liu, A. G. Schwing, K. Kundu, R. Urtasun, and S. Fidler, "Rent3d: Floor-plan priors for monocular layout estimation," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 3413–3421, 2015.

[31] Q. Hou, M.-M. Cheng, X. Hu, A. Borji, Z. Tu, and P. H. Torr, "Deeply supervised salient object detection with short connections," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 41, no. 4, pp. 815–828, 2019.

[32] O. Russakovsky *et al.*, "Imagenet large scale visual recognition challenge," *Int. Journal of Computer Vision*, vol. 115, pp. 211–252, 2015.

[33] H. Xie *et al.*, "Autonomous multi-robot navigation and cooperative mapping in partially unknown environments," *IEEE Trans. on Instrumentation and Measurement*, vol. 72, pp. 1–12.

[34] Y. Xiu *et al.*, "Finite-time sideslip differentiator-based los guidance for robust path following of snake robots," *IEEE/CAA Journal of Automatica Sinica*, vol. 10, no. 1, pp. 239–253, 2023.

[35] J. Zhang *et al.*, "Using tabu search to avoid concave obstacles for source location," *IEEE Trans. on Intelligent Transportation Systems*, 2023.

[36] J. Zhang, Y. Lu, L. Che, and M. Zhou, "Moving-distance-minimized pso for mobile robot swarm," *IEEE Trans. on Cybernetics*, vol. 52, no. 9, pp. 9871–9881, 2021.

[37] S. Han, K. Zhu, M. Zhou, and X. Liu, "Joint deployment optimization and flight trajectory planning for uav assisted iot data collection: A bilevel optimization approach," *IEEE Trans. on Intelligent Transportation Systems*, vol. 23, no. 11, pp. 21492–21504, 2022.

[38] H. Huang, W. He, Q. Fu, X. He, and C. Sun, "A bio-inspired flapping-wing robot with cambered wings and its application in autonomous airdrop," *IEEE/CAA Journal of Automatica Sinica*, vol. 9, no. 12, pp. 2138–2150, 2022.

[39] Z. Cao, J. Li, D. Zhang, M. Zhou, and A. Abusorrah, "A multi-object tracking algorithm with center-based feature extraction and occlusion handling," *IEEE Trans. on Intelligent Transportation Systems*, vol. 24, no. 4, pp. 4464–4473, 2022.