

Article

High-Precision BEV-Based Road Recognition Method for Warehouse AMR Based on IndoorPathNet and Transfer Learning

Tianwei Zhang ¹, Ci He ^{1,2,*}, Shiwen Li ¹, Rong Lai ¹, Zili Wang ³ , Lemiao Qiu ³  and Shuyou Zhang ³

¹ School of Mechanical Engineering, Yangzhou University, Yangzhou 225127, China; mz120210803@stu.yzu.edu.cn (T.Z.); mz120210825@stu.yzu.edu.cn (S.L.); mz120220865@stu.yzu.edu.cn (R.L.)

² Canny Elevator, Suzhou 215213, China

³ Department of Mechanical Engineering, Zhejiang University, Hangzhou 310058, China; ziliwang@zju.edu.cn (Z.W.); qiulm@zju.edu.cn (L.Q.); zsy@zju.edu.cn (S.Z.)

* Correspondence: heci_yzu@yzu.edu.cn; Tel.: +86-15267010098

Featured Application: An autonomous mobile robot (AMR) is a self-navigating vehicle capable of executing specific tasks such as transportation and cruising. As a crucial component of the logistics system in Industry 4.0 and smart factories, AMRs leverage sensors, including cameras, to gather information about their external environment and chart navigational paths. Their applications are diverse, spanning across warehousing, manufacturing, healthcare, hazardous locations, and specialized industries. Road recognition, a key feature of AMRs, employs technologies like semantic segmentation to extract high-level scene information, including object locations and semantic data from images. This technology can be integrated with Visual Simultaneous Localization and Mapping (SLAM), enhancing the AMR's ability to navigate complex environments. This integration is pivotal in enabling AMRs to successfully complete their navigation tasks.

Abstract: The rapid development and application of AMRs is important for Industry 4.0 and smart logistics. For large-scale dynamic flat warehouses, vision-based road recognition amidst complex obstacles is paramount for improving navigation efficiency and flexibility, while avoiding frequent manual settings. However, current mainstream road recognition methods face significant challenges of unsatisfactory accuracy and efficiency, as well as the lack of a large-scale high-quality dataset. To address this, this paper introduces IndoorPathNet, a transfer-learning-based Bird's Eye View (BEV) indoor path segmentation network that furnishes directional guidance to AMRs through real-time segmented indoor pathway maps. IndoorPathNet employs a lightweight U-shaped architecture integrated with spatial self-attention mechanisms to augment the speed and accuracy of indoor pathway segmentation. Moreover, it surmounts the challenge of training posed by the scarcity of publicly available semantic datasets for warehouses through the strategic employment of transfer learning. Comparative experiments conducted between IndoorPathNet and four other lightweight models on the Urban Aerial Vehicle Image Dataset (UAVID) yielded a maximum Intersection Over Union (IOU) of 82.2%. On the Warehouse Indoor Path Dataset, the maximum IOU attained was 98.4% while achieving a processing speed of 9.81 frames per second (FPS) with a 1024 × 1024 input on a single 3060 GPU.

Keywords: road recognition; semantic segmentation; transfer learning



Citation: Zhang, T.; He, C.; Li, S.; Lai, R.; Wang, Z.; Qiu, L.; Zhang, S. High-Precision BEV-Based Road Recognition Method for Warehouse AMR Based on IndoorPathNet and Transfer Learning. *Appl. Sci.* **2024**, *14*, 4587. <https://doi.org/10.3390/app14114587>

Academic Editor: Stéfano Frizzo Stefenon

Received: 5 May 2024
Revised: 18 May 2024
Accepted: 20 May 2024
Published: 27 May 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

In Industry 4.0 and smart factory evolution, companies are developing from fully automated assembly lines to complete automation encompassing the entire production floor to the warehouse [1]. Automated Guided Vehicles (AGVs), first introduced in 1955, are developed as high-automated, low-labor logistic platforms to support large-scale material flows in production plants, warehouses, and distribution centers. Replacing conventional

trucks with AGV systems greatly benefits the sustainable development of logistic operations [2] because of the safety, efficiency, economy, accuracy and productivity [3]. Industry 4.0 signifies a fresh industrial framework catalyzed by disruptive technologies, with the potential to revolutionize manufacturing into a cyber-physical system seamlessly uniting products, people, and processes. Ilias P. Vlachos et al. opt for AGVs and the Internet of Things (IoT) as focal points to devise an operational strategy aiding managers in incorporating Industry 4.0 technologies into their manufacturing systems, thereby realizing lean automation objectives [4].

However, the adaptation challenge of AGVs becomes particularly prominent [5]. Frequent changes take place in large-scale flat warehouse environments, which make accurate prediction and description difficult to obtain, especially in multi-variety, small-batch, and flexible manufacturing production plants. The dense arrangement of finished products, raw materials, and assorted equipment presents significant hurdles. The frequent movement of goods in and out of storage results in frequent shifts in obstacle positions, bringing inconvenience and even danger to navigation based on predefined static maps [6]. In a factory assembly line environment, Vehicle-to-Everything (V2X) technology, based on IoT engineering, allows AMRs to interact with each other and with workstation units to efficiently complete scheduling and production tasks. However, V2X technology encounters significant challenges in achieving environmental perception in warehouse settings, where most obstacles are unable to exchange information [7]. Furthermore, the limited sensing range of onboard sensors restricts the effectiveness of environment perception and navigation [8]. The lack of comprehensive global directional guidance may cause AGVs to struggle to anticipate pathway changes, potentially leading to deviations caused by circumnavigation errors or even loss of orientation [9].

In order to solve the problem, AGV guidance systems have progressed through various stages including mechanical, optical, induction, inertial, and laser guidance, culminating in today's vision-based systems, which have promoted the traditional AGV to AMR [10]. The advantage of visual-based systems lies in the ability to enhance adaptability to various environments through diverse algorithms to ensure competitiveness [11], compared with traditional methods such as laying magnetic strips on the ground and detecting obstacles using 2D and 3D radars [12]. NASA and the Jet Propulsion Laboratory's autonomous robots rely on IMU-based visual-inertial odometry and precomputed trajectory planning for navigating static obstacle courses [6]. Ubiquitous sensors, powerful onboard computers, artificial intelligence (AI) methods, and SLAM technologies are utilized to perceive the environment and navigate within facilities without the need for predefined and implemented reference points [13].

Road recognition plays a pivotal role in geographic information processing and computer vision, with the primary objective of accurately identifying and extracting road information from visual images or map data [14]. In outdoor road applications, the convergence of multiple LiDAR devices may create a high-energy laser mist. Although it has not yet been experimentally proven whether prolonged or repeated exposure to this laser mist is harmful to the human retina, scholars still prefer to prioritize the development of vision systems for these scenarios [15]. Early researchers exploring computer vision for mobile robot navigation proposed methods like a One-Versus-Rest (OVR)-type Support Vector Machine (SVM) classifier with relaxed constraints and kernel functions [16]. This approach integrated illumination-aware image segmentation, and guided color enhancement and adaptive threshold segmentation to tackle path recognition challenges in complex lighting conditions. By employing particle filters, researchers have successfully navigated the fusion of data from conventional one-dimensional image databases and directed graphs. This approach facilitates precise location identification within networked environments, including urban streets and multi-lane roads where lateral platform shifts occur [17]. As instance segmentation techniques represented by neural networks have gained prominence [18], road recognition has evolved from basic filter-based region recognition to pixel-level semantic segmentation relying on Convolutional Neural Networks (CNNs).

The rapid development of deep learning techniques, notably CNNs, has greatly advanced the effectiveness and efficiency of road image recognition. Semantic segmentation technology enhances the precision of road recognition by effectively distinguishing road areas from other objects, thereby improving the accuracy of road extraction and topological analysis. The advent of Fully Convolutional Networks (FCNs) [19] expanded image classification to the pixel level, mitigating the image reduction issue induced by pooling while enriching semantic information in features. However, traditional FCN approaches often overlook spatial pixel relationships, yielding fragmented segmentation outcomes [20]. Some researchers have utilized ResNet-18 with onboard cameras to differentiate between ground and walls [21]. The U-Net architecture, initially tailored for medical image segmentation, adopts a distinctive “U”-shaped configuration, with the encoder progressively reducing spatial dimensions and the decoder systematically restoring details and spatial attributes [22]. This unique design fortifies feature propagation and utilization efficacy, thereby enhancing segmentation precision and versatility. The encoder’s hierarchical arrangement ensures the extraction of semantic insights across multiple scales, facilitating comprehensive interpretation of image content [23]. Liu Yizhi et al. proposed a road intersection recognition method that combines a classification model based on GPS data with clustering algorithms [24]. They also introduced a technique utilizing real-time video captured by drones or pre-downloaded satellite multispectral images for road extraction and road condition detection [25]. Chenyang Feng et al. proposed a multi-path semantic segmentation network incorporating convolutional attention guidance to assist vehicles in road recognition [26].

Transfer learning provides an efficient manner of learning and transferring knowledge between previous tasks and the target task when the target data are rare, inaccessible, expensive and difficult to compile [27]. Transfer learning often involves three main strategies: direct transfer, feature extraction, and fine-tuning [28]. Direct transfer preserves the pre-trained model’s convolutional layers, replacing only the fully connected layers for the new task. Feature extraction uses the pre-trained model’s convolutional layers to extract feature vectors, which are then input into a new fully connected network. Fine-tuning adjusts selected convolutional layer parameters to better fit the new task, often freezing certain layers or training custom, fully connected layers for improved performance. Gianluca Fontanesi et al. [29] developed a Deep Q-Network (DQN) Lyapunov-based method for nonconvex Urban Aerial Vehicle (UAV) path design in various simulations and introduced a transfer learning strategy to enhance learning in new mmWave domains using knowledge from a sub-6 GHz domain. Piotr Mirowski et al. [30] proposed a deep reinforcement learning algorithm that helps models achieve navigation capabilities between cities without maps, realizing the transfer learning of task content. Given the scarcity of semantic data tailored for warehousing tasks in public datasets, this paper employs transfer learning to address the challenge of insufficient training samples and features.

However, traditional and deep-learning-based road recognition methods currently face significant challenges, including the misidentification of road surfaces and backgrounds. Environments with complex obstructions—such as buildings, trees, pedestrians, and vehicles—often lead to substantial misdetection issues. Additionally, roads that are narrow or short in length are prone to being overlooked, adversely affecting the overall efficiency, accuracy, and stability of these systems. Compounding these issues is the scarcity and high cost of high-quality datasets for large-scale indoor warehousing environments, which hampers the smooth implementation of neural-network-based road recognition. Fuentes-Pacheco et al. [31] surveyed VSLAM methods and found common issues such as efficiency degradation, unstable camera motion, and occlusions. These challenges are particularly prominent in scenarios involving moving objects, loop closure, robot kidnapping, and large-scale mapping. Angladon et al. [32] assessed and contrasted different real-time RGB-D visual odometry algorithms tailored for mobile devices, highlighting that only premium devices such as the iPad Air could uphold a satisfactory frame rate at a standard Video Graphics Array resolution.

Thus, there is still a need to improve the road recognition method to avoid the above problem. This paper proposes an IndoorPathNet path extraction scheme based on BEV images sourced from warehouse surveillance, aiming to tackle the challenge of intelligent road recognition for AMRs in vast and dynamic warehouse settings. The main contributions of the paper can be summarized as follows:

- (1) The proposed IndoorPathNet mitigates overfitting through structural optimization, and enhances its capacity to handle crucial information by integrating spatial attention mechanisms and dilated convolutions to widen the receptive field, thereby improving the efficiency and completeness of path feature extraction;
- (2) Pre-training implemented on the UAVID dataset facilitates the rapid learning of road features, augmenting its versatility and convergence speed;
- (3) To enhance generalization capability and training efficiency, this paper incorporates edge detection as an auxiliary task and employs random image cropping and stitching for data augmentation, thereby substantially enlarging the training set and enhancing image diversity.

The remaining sections of the paper are structured as follows: Section 2 provides a detailed exposition of IndoorPathNet's structural optimization, spatial self-attention mechanism, dataset, and transfer learning strategy; Section 3 showcases and discusses the comparative experiments between IndoorPathNet and four other models, along with the ablation experiments conducted on the Warehouse Indoor Path Dataset (WIPD); and finally, Section 4 provides conclusions and the future directions, respectively.

2. Materials and Methods

This section will offer an exhaustive elucidation of our model architecture, attention mechanism, dataset, and training approach. IndoorPathNet is structured around a U-shaped encoder–decoder framework, employing dilated convolutions to enlarge the receptive field and integrating spatial self-attention mechanisms to consolidate contextual semantic information. Section 2.1 will delineate the comprehensive architecture of IndoorPathNet. Section 2.2 will delve into the intricacies of the spatial self-attention mechanism. Section 2.3 briefly outlines the workflow in the application scenario. Section 2.4 will provide insights into the dataset utilized in this study. Finally, Section 2.5 will expound upon the training strategy grounded in the feature transfer.

2.1. Architecture of IndoorPathNet

As illustrated in Figure 1, IndoorPathNet represents an end-to-end semantic segmentation network tailored for extracting drivable paths within indoor warehouse environments. Its architecture comprises the following:

1. U-shaped Encoder–Decoder Structure: The original image is processed through encoding and decoding stages, facilitating the automatic segmentation and extraction of drivable paths from indoor BEV images;
2. Encoding Phase: In the encoding component, IndoorPathNet adopts a contraction path akin to U-Net, employing dilated convolution kernels with dilation factors of 2 and 5 in lieu of standard 3×3 convolution kernels; this integration, coupled with spatial self-attention mechanisms, serves to preserve critical spatial information while downsizing the feature map dimensions;
3. Decoding Phase: The decoding segment follows an expansion path reminiscent of U-Net, utilizing upsampling and feature enrichment based on attention mechanisms; it harnesses an enhanced spatial self-attention mechanism to prioritize the processing and reconstruction of regions deemed more crucial for prediction.

IndoorPathNet draws its design inspiration from U-Net, where the encoder's initial layer utilizes small convolution kernels (typically 3×3) to capture fundamental image features like edges and textures. This approach reduces model parameters while maintaining sensitivity to details. As the encoder progresses through subsequent layers, it transitions to extracting more abstract features such as image structures and shapes. This hierarchical

feature extraction ensures the network's ability to interpret image content across multiple scales, providing the semantic information necessary for precise segmentation. At the deepest level of U-Net, the fourth layer, along with its bottom bridge connections, performs the highest-level semantic feature extraction.

Expanding on this theoretical foundation, we conducted structural optimization based on the U-Net architecture, retaining only the upper three layers of the encoder–decoder structure for training. However, we encountered challenges as the model gradually began to overfit and struggled to differentiate between textures resembling walls and floors. Consequently, we reintroduced a bottleneck layer to introduce deeper features. This was coupled with spatial self-attention mechanisms and dilated convolutions, resulting in the development of the complete IndoorPathNet model.

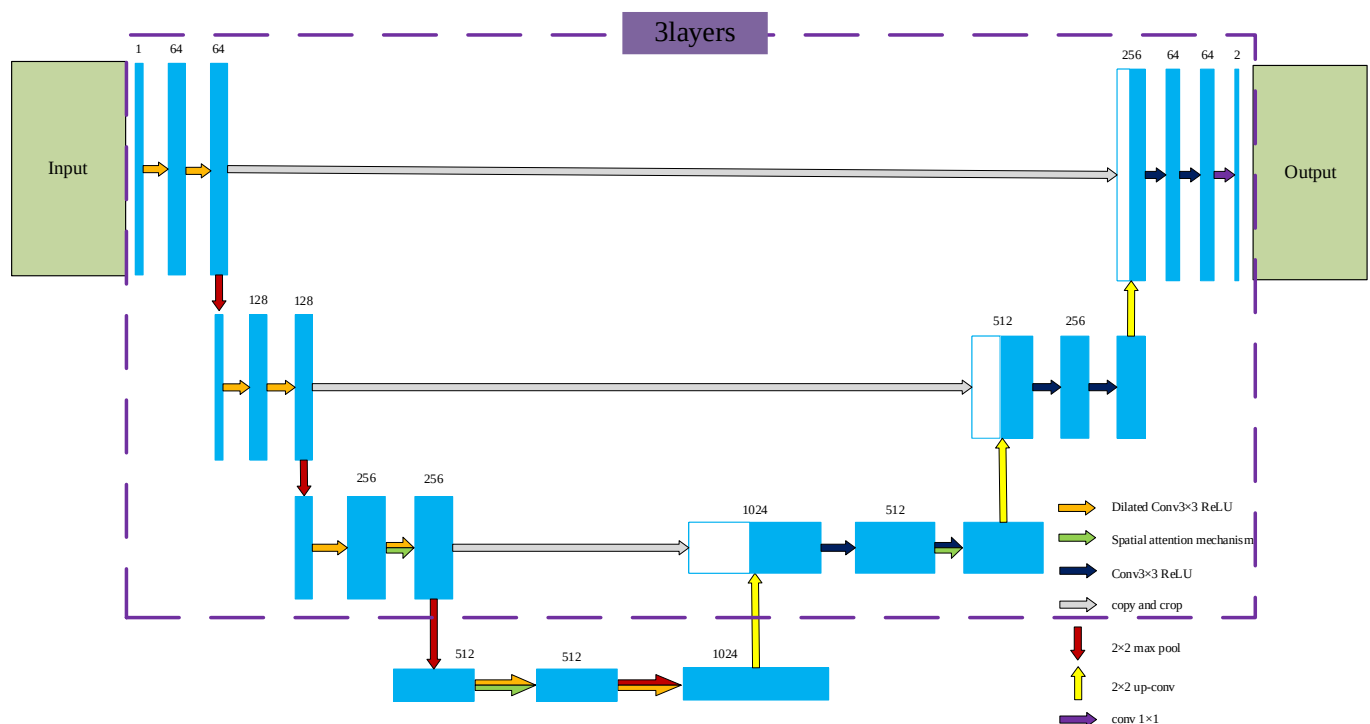


Figure 1. Architecture of the proposed IndoorPathNet.

2.2. Spatial Self-Attention Mechanism

The attention mechanism is crucial for managing information overload by prioritizing essential information with limited computational resources. It selectively emphasizes key data (V) by filtering out less pertinent details from a larger dataset (K), calculating weight coefficients to highlight more significant information. This feature enables the model to focus on areas critical for path extraction, thus optimizing the feature extraction process and enhancing model performance.

This study introduces a spatial self-attention mechanism to improve model performance further. To keep the network lightweight and reduce the number of parameters, this paper eschews traditional spatial and channel attention designs. Instead, it implements a modified spatial self-attention mechanism derived from a non-local self-attention framework, specifically tailored for binary classification problems, as shown in Figure 2.

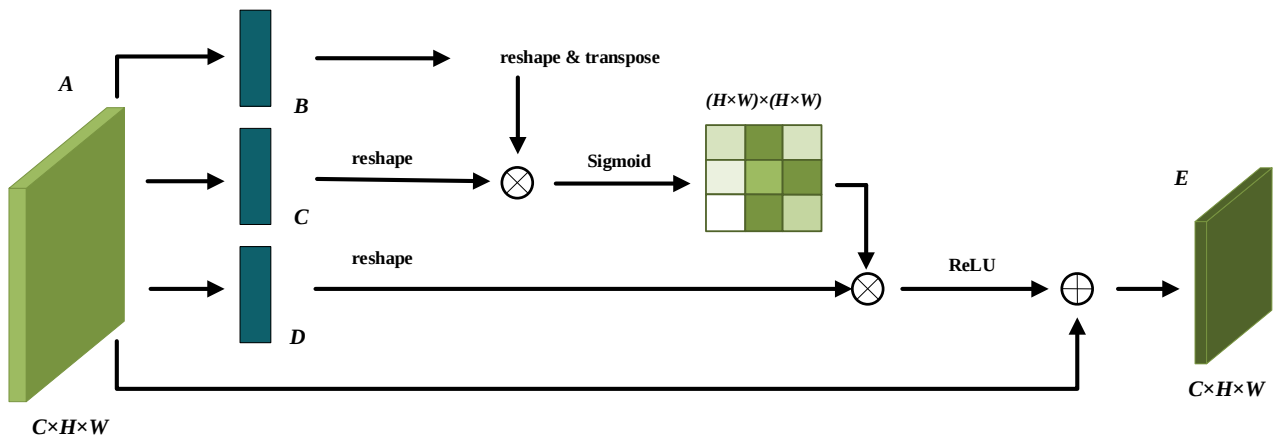


Figure 2. The spatial self-attention mechanism.

The process begins with a local feature A in $R^{C \times H \times W}$. This feature is input into a convolution layer, which produces two new feature maps, B and C , both of which are in $R^{C \times H \times W}$. These maps are then reshaped to $R^{C \times N}$, where $N = H \times W$ denotes the total number of pixels. Following this, matrix multiplication is executed between B and the transpose of C . A sigmoid layer is subsequently applied to calculate the spatial attention map, resulting in S in $R^{N \times N}$, as specified in Equation (1). This structured approach facilitates focused processing based on the significance of spatial relationships in the data.

$$S_{ji} = \frac{1}{1 + \exp(B_i \cdot C_j)} \quad (1)$$

The element S_{ji} quantifies the influence of the i -th position on the j -th position, with their correlation increasing as their feature representations become more similar. In this process, feature A is fed into a convolution layer to generate a new feature map $D \in R^{C \times H \times W}$, which is subsequently reshaped to $R^{C \times N}$. Following this, a matrix multiplication is conducted between D and the transpose of S , and the resulting matrix is reshaped back to $R^{C \times H \times W}$. Finally, this reshaped matrix is scaled by a parameter α and combined element-wise with the original feature A to yield the final output $E \in R^{C \times H \times W}$, as described in Equation (2). This method efficiently integrates contextual information, leveraging the spatial attention map to prioritize crucial areas for enhanced model performance.

$$E_j = \text{ReLU}(x) + A_j = \max(0, \alpha \sum_{i=1}^N (s_{ji} D_i)) + A_j \quad (2)$$

The scaling parameter is initially set to zero and gradually adjusted to assign increasing weight. From Equation (2), it is inferred that the feature at each position is a weighted sum of all positional features combined with the original feature. This configuration allows for a global contextual perspective, enabling the selective aggregation of contextual information guided by the spatial attention map.

2.3. Workflow in Application

Regarding our application scenario, this section provides a brief workflow introduction to help readers better understand the practical significance of the IndoorPathNet model in low-cost AMR warehouse navigation. The equipment utilized comprises a BEV camera, an AMR constructed on the Robot Operating System (ROS), and a computer equipped with a 3060 graphics card. Illustrated in the Figure 3, the workflow roughly encompasses four steps.

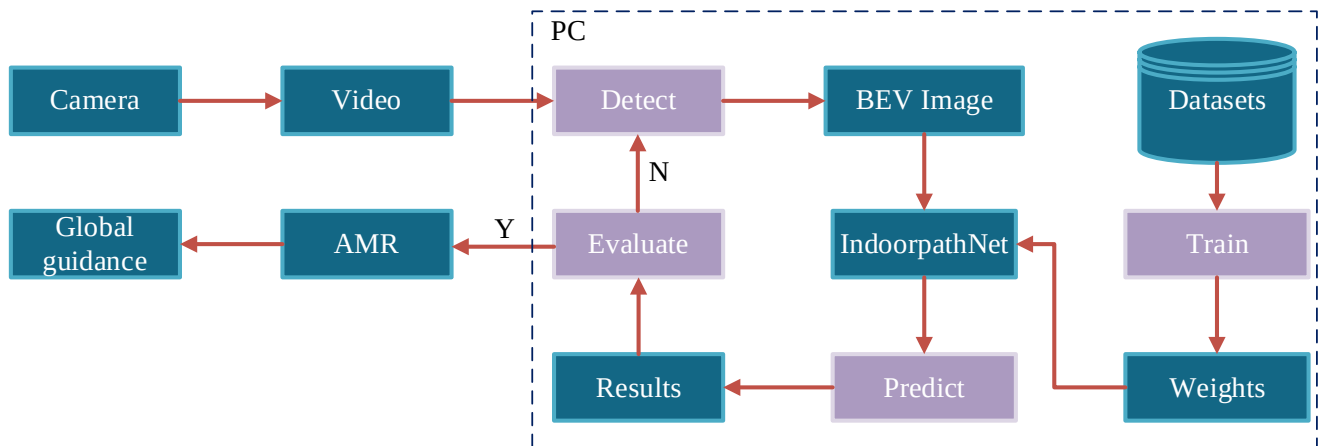


Figure 3. A high-level overview of the workflow in the application scenario.

- (1) After calibrating the camera to correct distortions [33], the signal is transmitted from the BEV camera to the PC via WIFI, followed by key frame extraction from real-time video based on a time interval strategy.
- (2) Images containing indoor paths are input into the pre-trained IndoorPathNet model for path segmentation.
- (3) Upon obtaining results, contour consistency is employed as an unsupervised evaluation metric to gauge the alignment of segmentation results' boundaries with clearly defined edges in the images. Additionally, consistency checks are conducted on multiple results within a short time frame.
- (4) Results passing the consistency check and achieving satisfactory evaluation scores are transmitted to the AMR end to provide global guidance. If results do not meet the criteria, the process is reverted to the first step to supplement key frame extraction for reevaluation.

After numerous experimental trials, the model proposed in this paper achieves an average processing speed of 0.126 s per 1024×1024 image. This filtering mechanism design guarantees that the AMR receives a global directional guidance every 0.5 s, ensuring consistent and timely navigation updates. These steps contribute to mitigating instances of low-quality path segmentation, or erroneous path recognition to some extent. The key frame extraction strategy, communication methods between devices, and evaluation methods for path recognition results can all be tailored to specific requirements. For instance, considering image quality in key frame selection can reduce recognition errors stemming from low image quality. Connecting AMRs and cameras to the Internet of Things network in smart factories via V2X protocols ensures rapid response and secure communication in emergencies. Additionally, histogram analysis can be incorporated to evaluate path segmentation results as necessary.

2.4. Overview of Dataset Used

This paper utilizes the following two datasets: the UAVID, as mentioned in the Abstract, which comprises street-view semantic data captured from drone perspectives, and the WIPD, a custom dataset specifically tailored for indoor path extraction tasks.

The UAVID is tailored for semantic segmentation tasks in urban scenes from UAVs, containing high-resolution aerial images from various urban settings with scenes including roads, buildings, and crowds. It consists of 30 video sequences with 300 images, densely labeled into 8 categories. To tailor it for path extraction, labels were refined by merging background and road categories into a binary classification for simplification. To aid the model in better edge detection, we performed secondary processing on the images using the Canny [34] algorithm.

The WIPD was developed using AI generation techniques to assess the semantic segmentation of warehouse indoor paths. As shown in the following results, this includes 50 images generated with the Midjourney model, depicting obstacles like goods and lifting equipment with prompts such as photorealistic, high-detail, BEV and cement floor. These images were manually annotated with LabelMe to differentiate drivable paths from backgrounds, resulting in JSON-formatted ground path labels.

To enhance the generalization capabilities of the model, we have implemented a novel data augmentation strategy for the WIPD dataset, specifically a random image cropping and stitching method. This approach involves using a Beta distribution to determine random center points for cropping four selected images. From these points, corresponding segments are extracted from each image and then seamlessly combined into a new, single image. This technique effectively expands the training dataset and boosts the model's ability to recognize and capture diverse features. Lastly, the paper implements zero-mean normalization to preprocess the input images. This involves normalizing all pixel values to the range of $[-1, 1]$ by subtracting the mean and dividing by the standard deviation for each channel. This preprocessing step is essential for enhancing the convergence speed of the IndoorPathNet semantic segmentation network.

2.5. Feature Transfer Learning Strategies

Consequently, this study utilizes UAVID as the source domain for transfer learning, pre-training the model to recognize path structure features. After this phase, the encoder's weights are frozen, and the model undergoes final training and fine-tuning on the target WIPD dataset. The complete transfer training workflow is depicted in Figure 4.

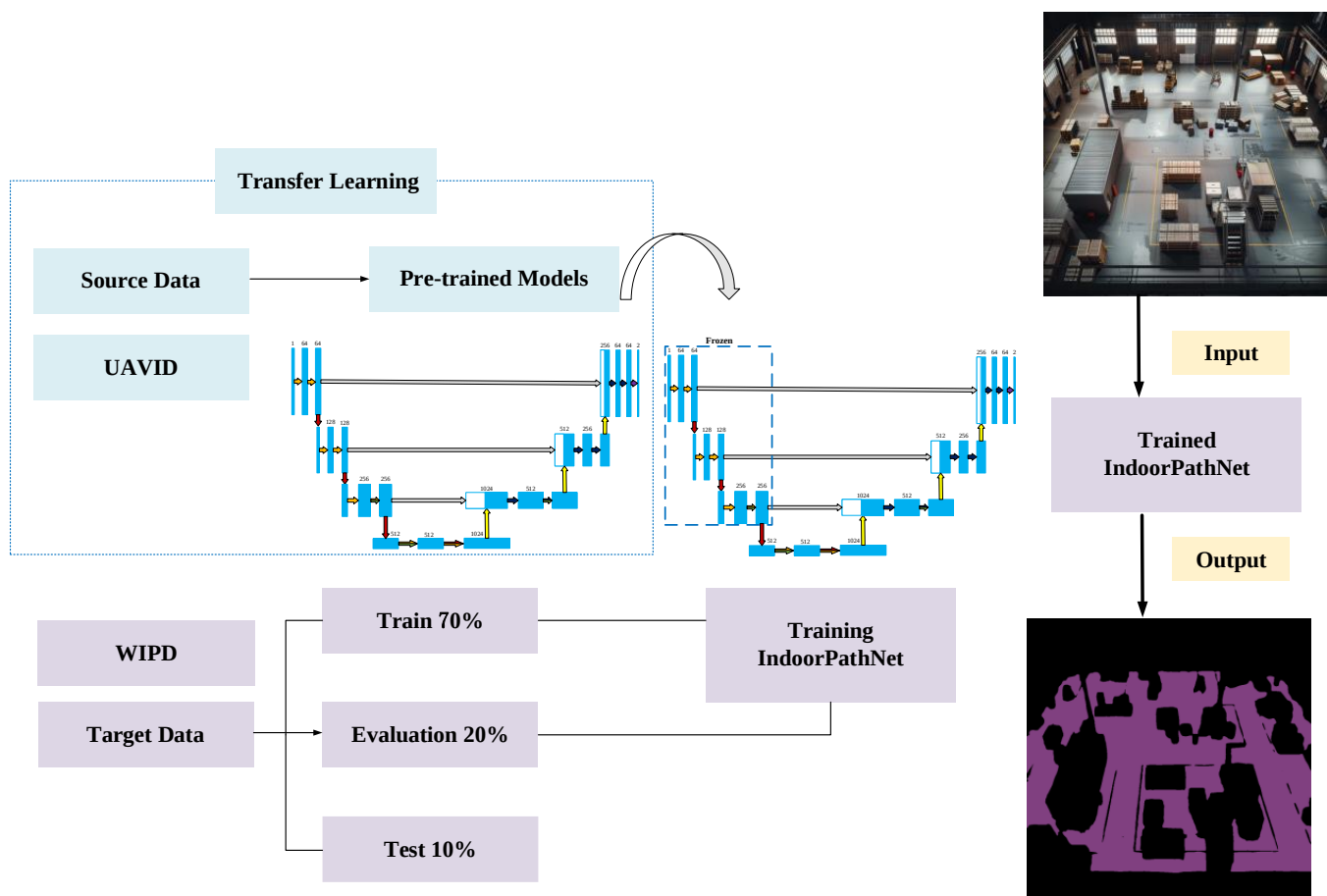


Figure 4. The overall schematic diagram of the feature transfer learning process.

In this study, the model employs a specific feature transfer training approach where, during the initialization phase of IndoorPathNet, the learning rates for some convolutional layers in the first three shallow layers of the pre-trained model are set to zero. This technique effectively preserves the distribution of features, such as textures and shapes, learned from the UAVID source domain data within the low-dimensional feature space. The model then undergoes further training with the WIPD dataset to significantly improve its ability to identify and extract path features in complex indoor environments. A diagram illustrating this process is shown in Figure 5.

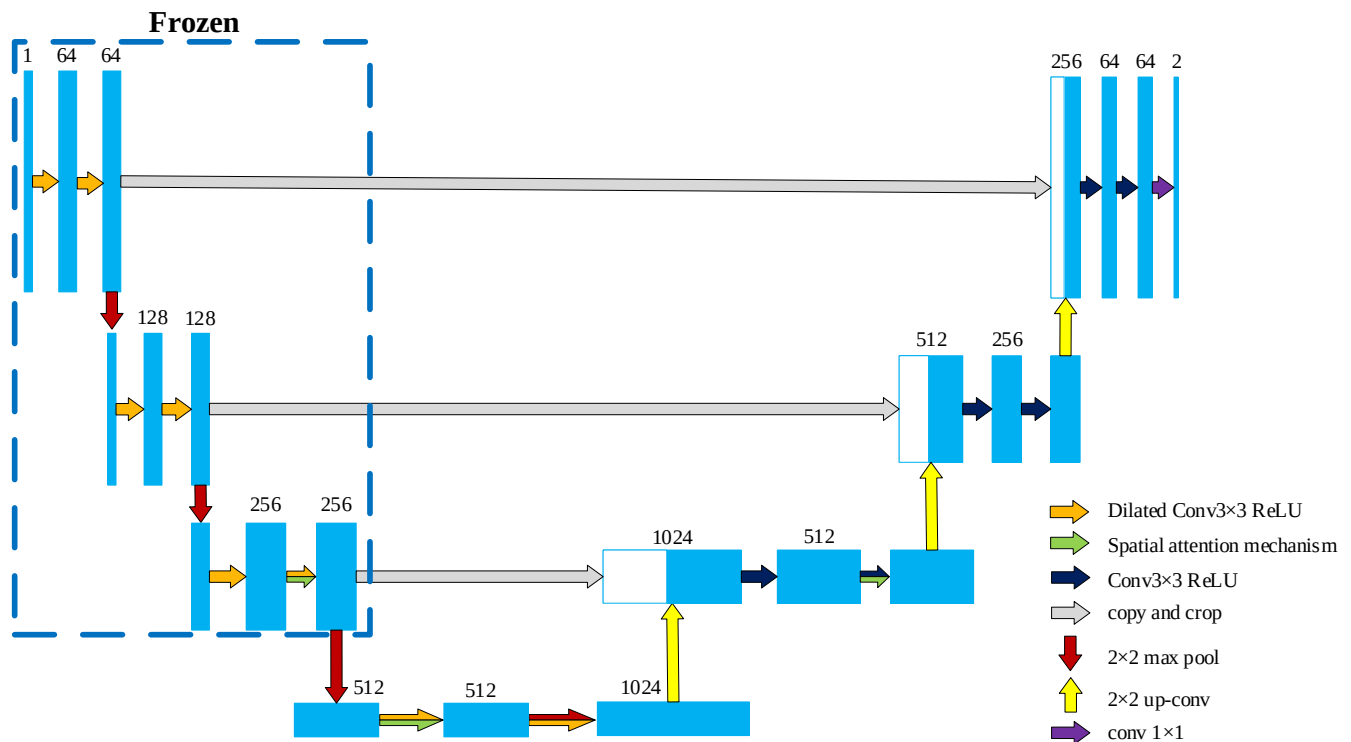


Figure 5. Encoder-frozen diagram of IndoorPathNet.

In the fine-tuning phase, this study uses the Adam optimizer for its dynamic learning rate adjustments based on gradient moments, enhancing both the training speed and model stability. The initial learning rates for the encoder's convolution modules are set at 1×10^{-4} , and 1×10^{-5} for the decoder's weights, tailored to the model's specifications. Mixed Precision Training (AMP) is adopted to improve efficiency under memory constraints. All parameters undergo L2 regularization with a 0.1 coefficient. The batch size is set to one to ensure precise data representation, with a maximum of 200 training iterations. To address pixel imbalance in indoor imagery, the model employs a combination of Dice loss [35] and cross-entropy loss [36], effectively optimizing the task of binary path segmentation.

3. Results

This section offers an analysis of the proposed method. Initially, we introduce the parameter metrics utilized for evaluating the model. Subsequently, we conduct comparative experiments between our model and four other lightweight semantic segmentation models using the same dataset, elucidating the distinctive advantages of our model through data comparisons and performance demonstrations. Finally, drawing from the ablation experiments, we delineate the improvement strategies and their efficacy in overcoming the challenges encountered during model design. In addition, we analyze the performance of the IndoorPathNet model in terms of accuracy and segmentation speed to emphasize its feasibility in warehouse-automated mobile robot navigation solutions.

3.1. Evaluation Metrics

This paper utilizes several metrics to assess the effectiveness of segmentation, including *Accuracy*, *Precision*, *Recall*, and *IoU*. A confusion matrix is employed initially to highlight the probability of misclassification among different categories. In the context of path extraction, paths are designated as positive and the background as negative. The evaluation begins by calculating the four basic indicators for binary classification tasks based on actual and model-predicted categories: True Positive (*TP*), True Negative (*TN*), False Positive (*FP*), and False Negative (*FN*). Below are the formulas used to compute these metrics:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)$$

$$Precision = \frac{TP}{TP + FP} \quad (4)$$

$$Recall = \frac{TP}{TP + FN} \quad (5)$$

$$IoU = \frac{TP}{TP + FP + FN} \quad (6)$$

3.2. Comparison Experiments

The experiment utilized three images from the UAVID dataset, characterized by paths with multi-scale features and obstructions from buildings, trees, pedestrians, and vehicles. The performance of the proposed methods was evaluated by comparison with four established network models—FCN-8s, FCN-32s, SegNet, and U-Net—through training on the same dataset. The results are visualized in Figure 6. The first row displays the original images, while the subsequent rows from the second to fifth show the outcomes for FCN-8s, FCN-32s, SegNet, and U-Net, respectively. The sixth row features the results from the IndoorPathNet model developed in this study. Areas highlighted in blue indicate where paths were incorrectly classified as background, which is categorized as *FN*. Areas in yellow show where the background was incorrectly classified as paths, which is documented as *FP*.

Notable performance disparities exist among FCN, its variants, and the proposed framework in the semantic segmentation task of distinguishing between paths and passable ground. As shown in Figure 6, it is revealed that FCN networks, particularly FCN-8s and FCN-32s, manifest the most pronounced misidentification issues. The improvement observed in FCN-32s over FCN-8s is marginal, often encountering both omission and commission errors in path identification. In contrast, the SegNet network exhibits some enhancement in reducing *FP* compared to the FCN network, but still struggles to prevent background misidentification as paths. Conversely, the U-Net network demonstrates significant improvement over the SegNet network, displaying a notable reduction in *FN* and a substantial decrease in *FP* compared to the FCN network. Although path continuity is enhanced, the detection outcomes remain incomplete. As indicated by the blue numbered boxes (d, e, f, and g) in the first image of different model effect diagrams, the issue of missing road detections improves with the increasing complexity of the models. This suggests that more complex models have a better capacity to capture detailed features. However, this has introduced a new problem, namely the misidentification of buildings becomes increasingly severe, peaking in the U-Net network effect diagram, which indicates that the model has overfitted. By comparing the yellow numbered boxes (1 to 6) between the U-Net network and IndoorPathNet, it can be observed that the misidentification of buildings has been reduced. Thus, all false negatives and false positives are actually concentrated on two issues: missing road detections and the incorrect identification of buildings. The IndoorPathNet proposed in this article achieves a good balance in addressing these two issues.

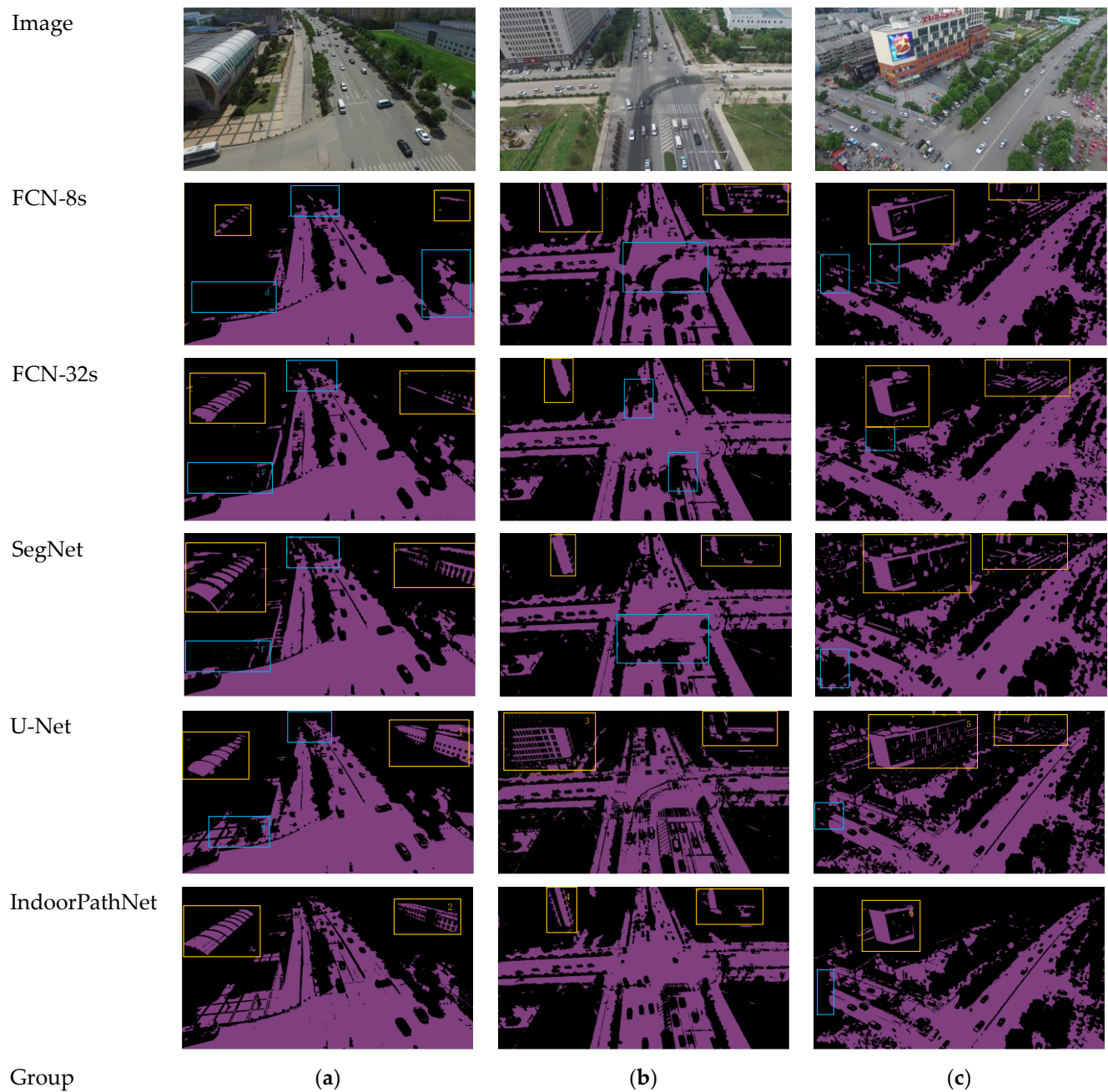


Figure 6. An intuitive comparison in the task scenario between current mainstream models and the pro-posed model; (a–c) represents different groups. The purple portion represents the segmented path. The rest is identified as the black background. The blue highlighted boxes represent *FN*, while the yellow ones represent *FP*.

Table 1 presents a detailed numerical comparison of the experiment. Among all models considered in the task scenarios, the proposed model outperforms the other four in terms of *precision*, *recall*, and *IoU* values. This superior performance underscores the significant advantages of our method over the other network models, particularly in enhancing the accurate identification of path targets, managing occlusion, and tackling discontinuity issues. The predictions generated by our model provide more comprehensive and precise path information. This capability allows for a more complete prediction of paths in small areas and short lengths within the images, thereby demonstrating a more robust path feature extraction ability compared to the other models.

Table 1. A numerical comparison in the task scenario between current mainstream models and the proposed model.

Group	(a)			(b)			(c)		
	P	R	IoU	P	R	IoU	P	R	IoU
FCN-8s	0.591	0.654	0.451	0.588	0.612	0.503	0.657	0.628	0.491
FCN-32s	0.553	0.611	0.409	0.529	0.574	0.480	0.563	0.623	0.510
SegNet	0.710	0.735	0.622	0.698	0.817	0.810	0.693	0.612	0.541
U-Net	0.822	0.895	0.762	0.872	0.890	0.759	0.610	0.533	0.632
IndoorPathNet	0.866	0.912	0.879	0.903	0.926	0.854	0.840	0.821	0.721

A quantitative comparison was conducted on the entire test set of the UAVID path dataset to further demonstrate the effectiveness of our method in path extraction. The quantitative results are shown in Table 2. Regarding precision metrics, the average improvement of our network is 36.3%, 40.3%, 13.8%, and 4.5% compared to FCN-8s, FCN-32s, SegNet, and U-Net networks, respectively. In terms of recall metrics, the average improvement of our network is 29.2%, 31.2%, 12.2%, and 5.6% compared to the four networks, respectively. Regarding the *IoU* metric, the average improvement of our network is 44.5%, 48.9%, 22.0%, and 6.1% compared to the four networks, respectively. Clearly, in terms of *precision*, *recall*, *F1 score*, and *IoU*, the quantitative results of our network outperform the other four network models.

Table 2. A comparison of whole UAVID between current mainstream models and the proposed model.

Model	P	R	IoU
FCN-8s	0.573	0.631	0.497
FCN-32s	0.602	0.681	0.503
SegNet	0.797	0.758	0.683
U-Net	0.813	0.802	0.764
IndoorPathNet	0.893	0.875	0.822

3.3. Ablation Experiments

To substantiate the contributions of structural optimization, attention mechanisms, data augmentation methods, dilated convolutions, and feature transfer methods to the outcomes of indoor path extraction, this chapter selects the image exhibiting the most severe overfitting phenomenon in the WIPD for ablation experiments. These experiments compare results before and after model enhancements based on the proposed improvement ideas. Additionally, effect maps, based on the U-Net model structure, are included for comparison to showcase structural optimization. Here, “3layers” denotes a variant of the U-Net structure after initial optimization, as illustrated in Figure 1.

The results of the ablation experiments reveal notable trends. In Figure 7c,d, the effects were unsatisfactory prior to the implementation of the feature transfer strategy. Following the addition of dilated convolutions, as depicted in Figure 7e, although the receptive field expanded, adjusting the model’s fitting direction remained challenging, resulting in misclassifications of walls as paths and the central area’s paths as background. Figure 7f,g illustrate that the spatial self-attention mechanism substantially enhanced path segmentation, enabling the segmentation of paths with stronger connectivity and intact edges. However, overfitting occurred, leading to misclassifications of walls and obstacle surfaces. The inclusion of data augmentation and other enhancements, as seen in Figure 7h, exacerbated the overfitting phenomenon of the U-Net model. Finally, the last image, Figure 7i, demonstrates the effectiveness of IndoorPathNet in extracting detailed and intact path semantic maps from BEV indoor storage images. Figure 8 shows the training metrics of IndoorPathNet on the WIPD.

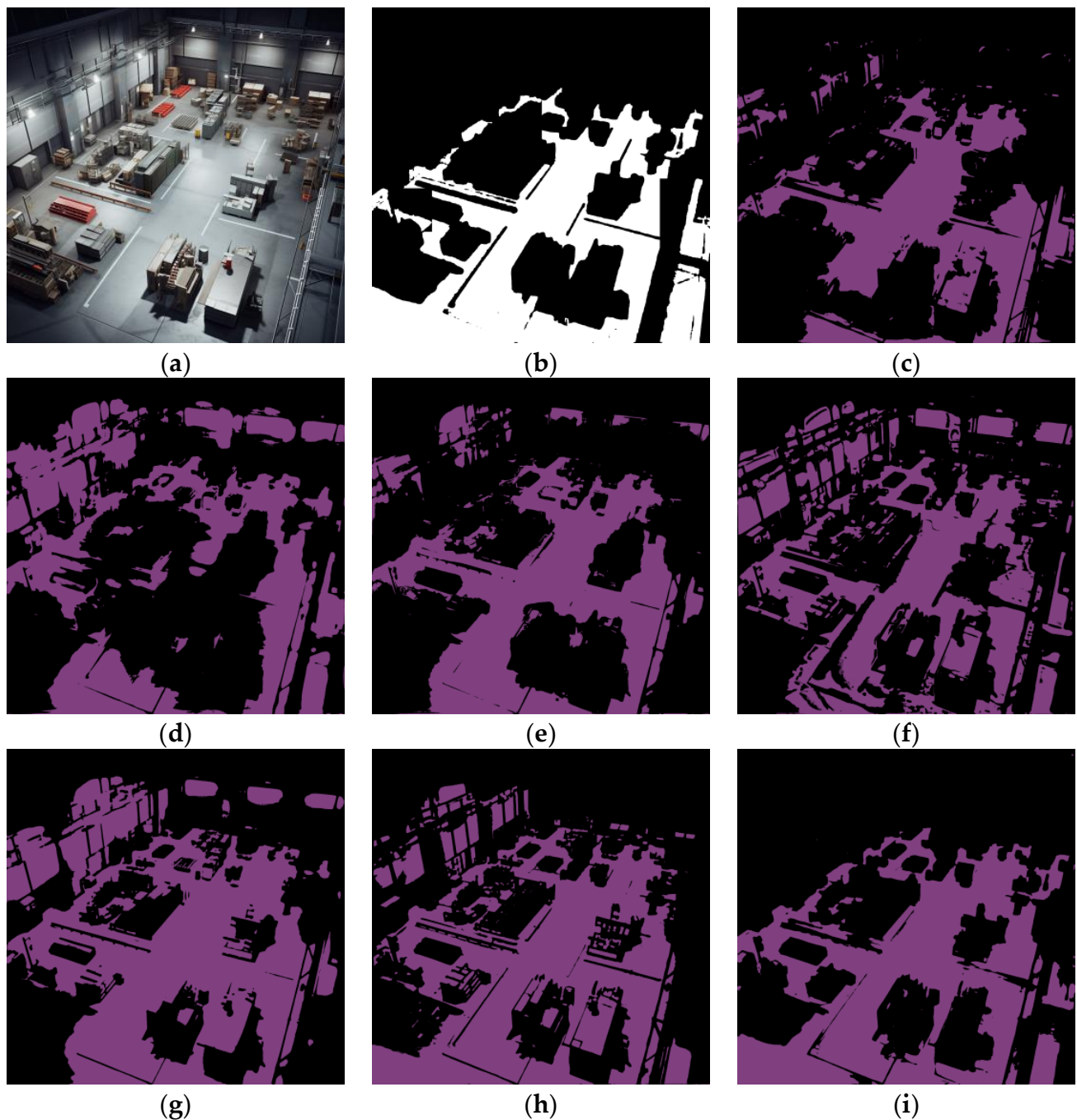


Figure 7. Results of ablation experiments in WIPD: (a,b) represent the image and mask, respectively; (c) represents the prediction of the three-layer model alone; (d) represents the prediction of U-Net alone; (e) represents three layers + dilated convolution + feature transfer; (f) represents three layers + dilated convolution + spatial self-attention mechanism + feature transfer; (g) represents U-Net + dilated convolution + spatial self-attention mechanism + feature transfer; (h) represents U-Net + dilated convolution + spatial self-attention mechanism + feature transfer + data augmentation; and (i) represents IndoorPathNet + feature transfer + data augmentation.

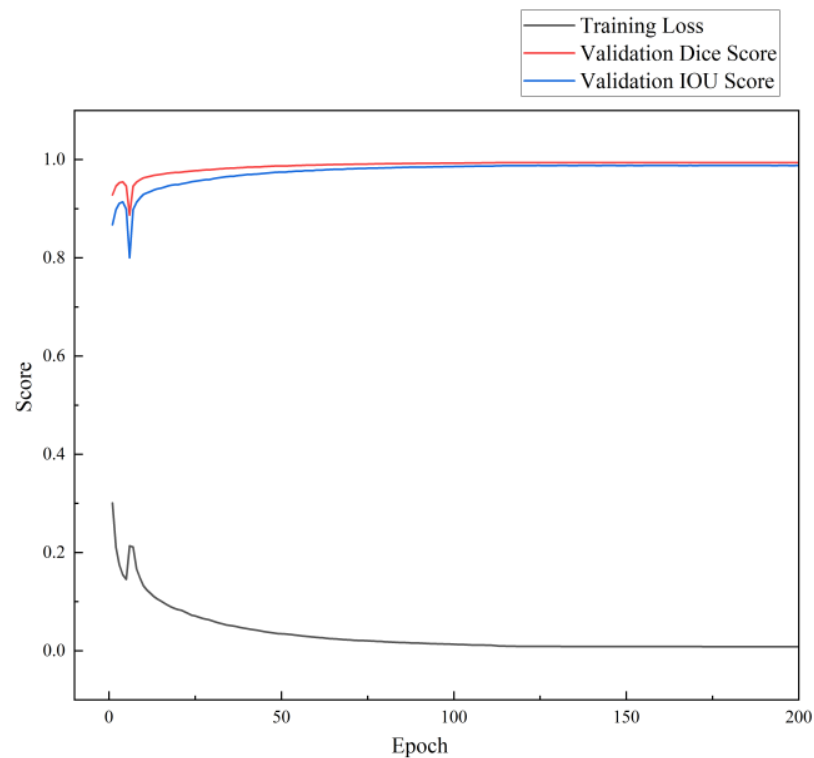


Figure 8. Training metrics on WIPD.

4. Conclusions

In this paper, a vision-based road recognition method based on an enhanced U-shaped IndoorPathNet and transfer learning is introduced, which is specifically designed for AMRs in large-scale flat warehouses. The efficiency and accuracy of road feature extraction is significantly improved, leveraging structural optimization, spatial self-attention mechanisms, dilated convolutions, feature transfer learning, edge detection auxiliary task and data augmentation. Compared with currently mainstream FCN-8s, FCN-16s, SegNet, and Unet, the proposed method exhibits exceptional performance on our proprietary WIPD, effectively providing real-time directional guidance for AMRs operating in large-scale dynamic warehouse environments. After extensive experimental trials, the model proposed in this paper achieves an average processing speed equivalent to approximately 9.81 frames per second (FPS) on a single 3060 GPU and an IOU of 98.4%. This performance fully satisfies the real-time global guidance needs of low-speed AMRs navigating within warehouse environments.

Although the IndoorPathNet is proposed to realize the path recognition task, some problems and shortcomings still arise in widespread practical use. The presented case studies are not conducted in natural light environments, anticipating occurrences where floor areas illuminated by light spots may be erroneously identified as obstacles, along with false detections of walls. In addition, path information of the entire flat warehouse is difficult to obtain based on a single BEV vision, which means that the possibility of utilizing multiple surveillance cameras to segment paths from diverse perspectives presents a promising solution to address perspective deficiencies. It is essential to emphasize that the semantic map of the paths generated by our method lacks physical scale and necessitates the acquisition of physical information through other sensors to meet the navigation requirements of AMRs. As a potential solution, local environmental perception by the onboard sensors of AMRs can effectively compensate for the absence of obstructed paths in the BEV, and enable fully automated scanning and mapping, which forms an important part of the multi-modal SLAM framework.

Author Contributions: Data curation, T.Z. and S.L.; Formal analysis, T.Z., C.H. and S.L.; Funding acquisition, C.H.; Investigation, T.Z. and C.H.; Methodology, T.Z.; Project administration, L.Q.; Resources, Z.W. and L.Q.; Software, T.Z. and R.L.; Supervision, S.Z.; Validation, C.H.; Visualization, T.Z.; Writing—original draft, T.Z.; Writing—review and editing, C.H. All authors have read and agreed to the published version of the manuscript.

Funding: This paper was funded by National Natural Science Foundation of China (Grant Number: 52305285), Natural Science Foundation of Jiangsu Province (Grant Number: BK20230592), Open Fund Project of State Key Laboratory of Fluid Power and Mechatronic Systems (Grant Number: GZKF-202326), Innovative Science and Technology Platform Project of Cooperation between Yangzhou City and Yangzhou University, China (Grant Number: YZ2020266).

Data Availability Statement: The data presented in this study are available on request from the first author, Tianwei Zhang. The data are not publicly available due to privacy.

Conflicts of Interest: Author Ci He was employed by the company Canny Elevator. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

- Adrodegari, F.; Bacchetti, A.; Pinto, R.; Pirola, F.; Zanardini, M. Engineer-to-order (ETO) production planning and control: An empirical framework for machinery-building companies. *Prod. Plan. Control* **2015**, *26*, 910–932. [\[CrossRef\]](#)
- Herrero-Perez, D.; Martinez-Barbera, H. Modeling Distributed Transportation Systems Composed of Flexible Automated Guided Vehicles in Flexible Manufacturing Systems. *IEEE Trans. Ind. Inform.* **2010**, *6*, 166–180. [\[CrossRef\]](#)
- Lass, S.; Gronau, N. A factory operating system for extending existing factories to Industry 4.0. *Comput. Ind.* **2020**, *115*, 103128. [\[CrossRef\]](#)
- Vlachos, I.P.; Pascazzi, R.M.; Zobolas, G.; Repoussis, P.; Giannakis, M. Lean manufacturing systems in the area of Industry 4.0: A lean automation plan of AGVs/IoT integration. *Prod. Plan. Control* **2023**, *34*, 345–358. [\[CrossRef\]](#)
- Schwesinger, D.; Spletzer, J. A 3D approach to infrastructure-free localization in large scale warehouse environments. In Proceedings of the 2016 IEEE International Conference on Automation Science and Engineering (CASE), Fort Worth, TX, USA, 21–25 August 2016; pp. 274–279.
- Yasuda, Y.D.V.; Martins, L.E.G.; Cappabianco, F.A.M. Autonomous visual navigation for mobile robots: A systematic literature review. *ACM Comput. Surv.* **2020**, *53*, 1–34. [\[CrossRef\]](#)
- Grosset, J.; Fougères, A.-J.; Djoko-Kouam, M.; Bonnin, J.-M. Multi-agent simulation of autonomous industrial vehicle fleets: Towards dynamic task allocation in V2X cooperation mode. *Integr. Comput. Aided Eng.* **2024**, *31*, 249–266. [\[CrossRef\]](#)
- Rea, A. *AMR System for Autonomous Indoor Navigation in Unknown Environments*; Politecnico di Torino: Torino, Italy, 2022.
- Meng, X.; Fang, X. A UGV Path Planning Algorithm Based on Improved A* with Improved Artificial Potential Field. *Electronics* **2024**, *13*, 972. [\[CrossRef\]](#)
- Fragapane, G.; de Koster, R.; Sgarbossa, F.; Strandhagen, J.O. Planning and control of autonomous mobile robots for intralogistics: Literature review and research agenda. *Eur. J. Oper. Res.* **2021**, *294*, 405–426. [\[CrossRef\]](#)
- Tong, Y.; Liu, H.; Zhang, Z. Advancements in Humanoid Robots: A Comprehensive Review and Future Prospects. *IEEE/CAA J. Autom. Sin.* **2024**, *11*, 301–328. [\[CrossRef\]](#)
- Wang, H.; Wang, B.; Liu, B.; Meng, X.; Yang, G. Pedestrian recognition and tracking using 3D LiDAR for autonomous vehicle. *Robot. Auton. Syst.* **2017**, *88*, 71–78. [\[CrossRef\]](#)
- Facchini, F.; Oleśków-Szłapka, J.; Ranieri, L.; Urbinati, A. A Maturity Model for Logistics 4.0: An Empirical Analysis and a Roadmap for Future Research. *Sustainability* **2019**, *12*, 86. [\[CrossRef\]](#)
- Xu, Y.; Xie, Z.; Feng, Y.; Chen, Z. Road extraction from high-resolution remote sensing imagery using deep learning. *Remote Sens.* **2018**, *10*, 1461. [\[CrossRef\]](#)
- Rateke, T.; Justen, K.A.; Chiarella, V.F.; Sobieranski, A.C.; Comunello, E.; Wangenheim, A.V. Passive Vision Region-Based Road Detection. *ACM Comput. Surv.* **2019**, *52*, 1–34. [\[CrossRef\]](#)
- Wu, X.; Sun, C.; Zou, T.; Li, L.; Wang, L.; Liu, H. SVM-based image partitioning for vision recognition of AGV guide paths under complex illumination conditions. *Robot. Comput. Integr. Manuf.* **2020**, *61*, 101856. [\[CrossRef\]](#)
- Pepperell, E.; Corke, P.; Milford, M. Routed roads: Probabilistic vision-based place recognition for changing conditions, split streets and varied viewpoints. *Int. J. Robot. Res.* **2016**, *35*, 1057–1179. [\[CrossRef\]](#)
- Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path aggregation network for instance segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8759–8768.
- Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
- Arthi, K.; Brintha, K. Segmentation and Extraction of Parcels from Satellite Images Using a U-Net CNN Model. *ISAR Int. J. Res. Eng. Technol.* **2024**, *9*.

21. Teso-Fz-Betoño, D.; Zulueta, E.; Sánchez-Chica, A.; Fernandez-Gamiz, U.; Saenz-Aguirre, A. Semantic Segmentation to Develop an Indoor Navigation System for an Autonomous Mobile Robot. *Mathematics* **2020**, *8*, 855. [\[CrossRef\]](#)
22. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, 5–9 October 2015; Proceedings, Part III 18. pp. 234–241.
23. Ding, H.; Jiang, X.; Shuai, B.; Liu, A.Q.; Wang, G. Semantic segmentation with context encoding and multi-path decoding. *IEEE Trans. Image Process.* **2020**, *29*, 3520–3533. [\[CrossRef\]](#)
24. Liu, Y.; Qing, R.; Zhao, Y.; Liao, Z. Road Intersection Recognition via Combining Classification Model and Clustering Algorithm Based on GPS Data. *ISPRS Int. J. Geo-Inf.* **2022**, *11*, 487. [\[CrossRef\]](#)
25. Xu, H.; Yu, M.; Zhou, F.; Yin, H. Segmenting Urban Scene Imagery in Real Time Using an Efficient UNet-like Transformer. *Appl. Sci.* **2024**, *14*, 1986. [\[CrossRef\]](#)
26. Feng, C.; Hu, S.; Zhang, Y. A Multi-Path Semantic Segmentation Network Based on Convolutional Attention Guidance. *Appl. Sci.* **2024**, *14*, 2024. [\[CrossRef\]](#)
27. Hosna, A.; Merry, E.; Gyalmo, J.; Alom, Z.; Aung, Z.; Azim, M.A. Transfer learning: A friendly introduction. *J. Big Data* **2022**, *9*, 102. [\[CrossRef\]](#) [\[PubMed\]](#)
28. Zhuang, F.; Qi, Z.; Duan, K.; Xi, D.; Zhu, Y.; Zhu, H.; Xiong, H.; He, Q. A comprehensive survey on transfer learning. *Proc. IEEE* **2020**, *109*, 43–76. [\[CrossRef\]](#)
29. Fontanesi, G.; Zhu, A.; Arvaneh, M.; Ahmadi, H. A Transfer Learning Approach for UAV Path Design with Connectivity Outage Constraint. *IEEE Internet Things J.* **2023**, *10*, 4998–5012. [\[CrossRef\]](#)
30. Mirowski, P.; Grimes, M.; Malinowski, M.; Hermann, K.M.; Anderson, K.; Teplyashin, D.; Simonyan, K.; Zisserman, A.; Hadsell, R. Learning to navigate in cities without a map. *arXiv* **2018**, arXiv:1804.00168.
31. Fuentes-Pacheco, J.; Ruiz-Ascencio, J.; Rendón-Mancha, J. Visual simultaneous localization and mapping: A survey. *Artif. Intell. Rev.* **2015**, *43*, 55–81. [\[CrossRef\]](#)
32. Angladon, V.; Gasparini, S.; Charvillat, V.; Pribanić, T.; Petković, T.; Đonlić, M.; Ahsan, B.; Buel, F. An evaluation of real-time RGB-D visual odometry algorithms on mobile devices. *J. Real-Time Image Process.* **2017**, *16*, 1643–1660. [\[CrossRef\]](#)
33. Tardif, J.-P.; Sturm, P.; Trudeau, M.; Roy, S.; Intelligence, M. Calibration of cameras with radially symmetric distortion. *IEEE Trans. Pattern Anal. Mach. Intell.* **2008**, *31*, 1552–1566. [\[CrossRef\]](#) [\[PubMed\]](#)
34. Canny, J. A computational approach to edge detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **1986**, *8*, 679–698. [\[CrossRef\]](#)
35. Sudre, C.H.; Li, W.; Vercauteren, T.; Ourselin, S.; Jorge Cardoso, M. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support; DLMIA ML-CDS 2017 2017*, Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2017; pp. 240–248.
36. Zhang, Z.; Sabuncu, M.R. Generalized cross entropy loss for training deep neural networks with noisy labels. *arXiv* **2018**, arXiv:1805.07836.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.