Journal of Global Information Management
Volume 32 • Issue 1

A Deep Learning-Based Robot Analysis Model for Semantic Context Capturing by Using Predictive Models in Public Management

Zixuan Li, School of Public Policy and Management, China University of Mining and Technology, China
Chengli Wang, School of Public Policy and Management, China University of Mining and Technology, China*

ABSTRACT

In the realm of robotics, the ability to comprehend intricate semantic contexts within diverse environments is paramount for autonomous decision-making and effective human-robot collaboration. This article delves into the realm of enhancing robotic semantic understanding through the fusion of deep learning techniques. This work presents a pioneering approach: integrating several neural network models to analyze robot images, thereby capturing nuanced environmental semantic contexts. The authors augment this analysis with predictive models, enabling the robot to adapt the changing contexts intelligently. Through rigorous experimentation, our model demonstrated a substantial 25% increase in accuracy when compared to conventional methods, showcasing its robustness in real-world applications. This research marks a significant stride toward imbuing robots with sophisticated visual comprehension, paving the way for more seamless human-robot interactions and a myriad of practical applications in the evolving landscape of robotics.

KEYWORDS

CNN, Deep Learning, Robotics, Semantic Context, U-Net

INTRODUCTION

The significance of semantic analysis of robot images (Elmquist et al., 2022) lies in enhancing the perceptual ability and intelligence level of robots in complex environments, enabling them to accurately comprehend the surrounding environment and make corresponding decisions and actions. Through image semantic analysis (Lu et al., 2021), robots can identify and understand objects, scenes, and contexts in images, thereby better perceiving the surrounding environment. This capability is crucial for robots in navigating and interacting in complex and uncontrollable environments, such as disaster rescue (Zhao et al., 2022), outdoor exploration (Shah et al., 2021), and unknown environment detection tasks (Singh et al., 2022). Utilizing semantic analysis of robot images can help robots understand human actions, expressions, and intentions, facilitating better interaction with humans.

DOI: 10.4018/JGIM.335900 *Corresponding Author

In the fields of service robots (Okafuji et al., 2022) and social robots (Li et al., 2023), semantic analysis of robot images can assist robots in understanding user needs, providing more personalized and intelligent services and enhancing the user experience. Through image semantic analysis, robots can recognize different objects and situations and make corresponding decisions and plans based on this information. This ability is essential for key tasks such as autonomous navigation, obstacle avoidance, and task planning, enabling robots to independently accomplish various tasks without the need for real-time human intervention.

In the industrial sector (Lin et al., 2022), semantic analysis of robot images can help robots automatically identify and detect product defects, thereby improving production quality and efficiency. Moreover, it can be applied in intelligent warehousing (Dai et al., 2021), logistics (Wang & Chen, 2021), and other fields to achieve automated and intelligent logistics management. In the medical field (Sun et al., 2021), semantic analysis of robot images can be used for tasks such as surgical assistance, disease diagnosis, and rehabilitation treatment, assisting doctors in improving diagnostic accuracy and surgical precision and providing better rehabilitation services for patients.

An important real-world application case is presented herein, where the outcomes of semantic analysis of robot images can aid autonomous vehicles in comprehending their surroundings. Through image semantic analysis, vehicles can recognize roads, traffic signs, pedestrians, and other vehicles, enabling intelligent driving decisions. This constitutes a significant application case, as the comprehension of environmental images and semantic analysis thereof stand as pivotal modules within autonomous vehicles, facilitating intelligent automated driving (Chen & Zhang, 2023; Du & Chen, 2023).

Deep learning plays a crucial role in the task of semantic analysis of images for robots. Through convolutional neural networks (CNNs) (Nagata et al., 2021), robots can efficiently recognize objects and scenes within images. Recurrent neural networks (RNNs) (Kong, 2020) and long short-term memory networks (LSTMs) (Gao et al., 2021) are employed to handle image sequences, aiding robots in understanding dynamic scenes and predicting object movements. The integration of CNNs and RNNs enables comprehensive modeling of both image content and contextual information. Transfer learning, utilizing pretrained models, enhances performance on small-scale datasets. Generative adversarial networks (GANs) (Kushwaha et al., 2022) are used to generate images relevant to specific tasks, simulating complex environments. Self-supervised learning and reinforcement learning methods assist robots in learning features from images and optimizing decision-making strategies. These deep-learning techniques enable robots to comprehend images rapidly and accurately, enhancing their perceptual abilities and intelligence levels in complex environments and allowing them to better address various challenges. The deep-learning models commonly used in the research on semantic analysis of images for robots are as follows:

(1) U-Net (Li et al., 2021), a classic fully convolutional neural network, incorporates an encoder–decoder architecture designed for image-segmentation tasks. In the realm of robot-assisted surgery, U-Net has application in the segmentation of organs within medical images, aiding robots in precise localization and manipulation.
(2) DeepLab (Zhenzhen et al., 2021) utilizes deep convolutional neural networks and atrous convolution techniques to effectively capture contextual information within images. In the domain of autonomous vehicles, DeepLab is employed for road segmentation, enabling vehicles to recognize road boundaries and obstacles.
(3) Mask R-CNN (Zhang et al., 2021) combines object detection and semantic segmentation, allowing simultaneous detection of object positions and generation of precise segmentation masks. Within robotic grasping tasks, Mask R-CNN is utilized for object detection and segmentation, enabling robots to accurately grasp objects of various shapes.
(4) FCN (Cen, 2023) represents an end-to-end fully convolutional network that maps input images pixel-wise to semantic segmentation maps. In the realm of unmanned aerial vehicle (UAV) image

processing, FCN is employed for land cover classification and segmentation, empowering robotic systems to comprehend diverse features on the ground.

(5) SegNet (Turgut et al., 2022), a lightweight convolutional neural network, is particularly suited for real-time image segmentation. In the domain of service robotics, SegNet is harnessed for environmental perception and navigation, assisting robots in obstacle avoidance and path recognition.

This study proposes a novel image semantic segmentation model that combines the efficient feature extraction capability of CNNs, the excellent semantic segmentation performance of the U-Net architecture, and an introduced multi-head attention mechanism (Ning et al., 2023). First, we utilize CNN networks to extract features from input images, capturing crucial information within the images. Subsequently, we employ the U-Net structure for image semantic segmentation, achieving precise segmentation of different objects and scenes within the images. To further enhance the model's performance, we introduce an attention multi-head mechanism; this enables the model to focus more on important regions within the images, thus improving segmentation accuracy and precision. By synergistically integrating these three key components, our designed model demonstrates significant improvements in image semantic segmentation tasks, offering enhanced representational capabilities and finer segmentation results. This research provides robust support for the field of robotic visual perception in both research and practical applications.

The three contribution points of this paper are as follows.

(1) Integration of three key components: the model is unique in that it incorporates the efficient feature extraction of CNNs and the efficient feature extraction capability of CNN models for real-time processing of robotic images.
(2) Introduction of multi-attention mechanism: the multi-attention mechanism enables the model to adaptively focus on important regions in the image. This mechanism enhances the robustness of the model and its adaptability to complex environments, in addition to flexibly adjusting the focusing area, thus improving the ability to process complex image scenes.
(3) Optimization of image semantic segmentation performance: the model incorporates the U-Net model for image semantic segmentation, which achieves a comprehensive optimization of robotic image semantic understanding.

RELATED WORK

Image Semantic Segmentation Context

Image semantic segmentation (Balachandran & Ranganathan, 2023), a pivotal technology in the field of computer vision, aims to assign each pixel in an image to a specific semantic category, enabling a detailed delineation of semantic information within the image. In the realm of robotics, there is a growing focus on research concerning image semantic segmentation. Through the utilization of deep-learning techniques, particularly deep convolutional neural networks, significant progress has been made. These models not only accurately differentiate various objects and scenes against complex backgrounds but also enhance a robot's understanding and perception of its environment. In the field of robotics, image semantic segmentation finds extensive applications in tasks such as autonomous navigation, environment modeling, object recognition, and grasping. These studies not only provide crucial support for intelligent decision-making and autonomous behavior in robots but also lay a solid foundation for the development of safer, more efficient, and intelligent robotic systems (Ye et al., 2023; Ye & Zhao, 2023).

The application of image semantic context analysis in the field of robotics exhibits significant advantages. First, it endows robots with precise perceptual capabilities, enabling them to accurately

comprehend and differentiate various objects and structures in the surrounding environment, thereby supporting intelligent decision-making and operations in complex scenarios. Second, image semantic context analysis provides crucial cues for autonomous navigation, allowing robots to recognize roads, obstacles, and other key features, facilitating safe and efficient autonomous movement. Moreover, this technology forms the foundation for enhanced interactivity, enabling robots to engage in more intelligent and human-like interactions with users.

However, evident challenges exist. Image semantic context analysis entails high computational complexity; it demands substantial hardware performance and computational resources, which could be limiting, especially in embedded systems and mobile robots. There is a substantial requirement for extensive, well-labeled, high-quality data for training deep-learning models, resulting in high costs associated with data acquisition and processing. Additionally, in dynamically changing and intricate environments, enhancing the algorithm's robustness and real-time performance is a critical area for improvement. Despite the immense potential of image semantic context analysis in the robotics domain, continuous innovation and breakthroughs are still required in aspects such as algorithmic performance, hardware adaptation, and data support (Liu & Chen, 2023).

Robot Environment-Sensing Technology

Robot environmental-perception technology refers to the ability of robots to utilize sensors, cameras, and other perception devices to collect, process, and analyze data, enabling real-time acquisition and comprehension of surrounding environmental information. This technology endows robots with intelligent sensory capabilities, allowing them to adapt to diverse and complex environments and respond accordingly. In modern robotic applications, environmental-perception technology plays a pivotal role. In the industrial sector, robots employ environmental-perception technology for smart manufacturing, enabling automated assembly and quality inspection. In the domain of service robots, environmental-perception technology enables robots to recognize and navigate around obstacles, ensuring safe interactions with human users (Ye & Zhao, 2023; Ye et al., 2023). In agriculture and mechanization, robots utilize environmental-perception technology to achieve smart agriculture, facilitating crop monitoring and automated cultivation. In the field of health care, robots' environmental-perception capabilities enable them to assist surgeons, enhancing surgical precision and safety. Overall, robot environmental-perception technology has extensive applications in various sectors, leading to positive impacts such as enhanced productivity, reduced labor intensity, and improved health-care services (Cong et al., 2021).

The application of robot environmental perception technology plays a crucial role in modern industrial and service sectors, exhibiting evident advantages. First, environmental-perception technology endows robots with high intelligence, enabling them to accurately recognize various features in the surrounding environment, facilitating autonomous navigation, obstacle avoidance, and natural interactions with human users. This not only enhances productivity but also improves service quality, propelling the development of industrial automation and intelligent services. Second, robot environmental-perception technology excels in handling hazardous environments, such as in tasks like fire rescue and nuclear-radiation cleanup, where robots can replace humans in dangerous operations, thereby enhancing safety.

However, challenges persist in the application of robot environmental-perception technology. The stability and robustness of the technology need further improvement, especially in complex and dynamic environmental conditions. The cost of hardware and algorithms remains relatively high, limiting widespread adoption. Moreover, issues related to data privacy and security need to be addressed, particularly in applications involving personal privacy in service-robot contexts. Therefore, while enhancing technological performance, further research on data security and privacy protection is essential to achieve a broader and safer application of robot environmental-perception technology (Chen & Zhang, 2023).

U-Net Model

The U-Net model is a classic fully convolutional neural network initially proposed by researchers in the field of medical-image segmentation. Its distinctive architecture consists of a contracting path (encoder) and an expansive path (decoder), enabling highly accurate semantic segmentation in image-segmentation tasks. In U-Net, information is first downsampled through multiple convolutional layers and then upsampled through deconvolution and skip connections, ultimately generating pixel-level segmentation results. This architecture allows U-Net to capture features at different scales, providing more accurate and detailed segmentation. In the field of robot image processing, the U-Net model has widespread applications in various tasks. For instance, in robot visual navigation, U-Net can be utilized for map construction and environment perception, assisting robots in recognizing obstacles, roads, and other crucial features. In robot-assisted surgery, U-Net is employed for organ segmentation in medical images, enabling precise localization and manipulation by the robot. U-Net is used in industrial robots for detecting and locating product components, enhancing the automation level of production lines. The U-Net model, owing to its outstanding performance and broad applicability, plays a significant role in robot image processing, providing substantial support for improving robots' perceptual and intelligent decision-making capabilities (Patel et al., 2021).

The U-Net model demonstrates significant advantages in robot image analysis. First, its unique encoder–decoder architecture efficiently captures local and global features, providing highly accurate semantic segmentation for precise robotic perception. Second, the use of skip connections enables the model to transmit information across different layers, allowing it to handle images with varying scales and complexities and enhancing segmentation robustness. Additionally, U-Net's end-to-end training capability reduces information loss between feature extraction and segmentation tasks, improving overall performance.

However, the U-Net model also exhibits certain limitations. Its performance heavily relies on the quality of large-scale annotated data, potentially affecting its generalization ability due to dataset constraints. U-Net may experience performance degradation when dealing with very large or very small objects, necessitating additional handling methods to address these issues. The complexity of the U-Net model demands significant computational resources, limiting its application in embedded systems or mobile robots. U-Net's robustness against image transformations such as rotation, scaling, and deformation is relatively weak, potentially impacting segmentation accuracy in certain real-world scenarios. In summary, while the U-Net model holds promising prospects for robot image analysis, further research and improvements are necessary to enhance its adaptability and robustness in various complex environments and tasks (Feng & Chen, 2022).

METHOD

Overview

In our research, we have developed an innovative methodology for advanced image analysis by integrating state-of-the-art deep-learning techniques. Our approach combines convolutional neural networks for robust feature extraction, U-Net architecture for intricate semantic context analysis, and the incorporation of a multi-head attention mechanism.

Our research leverages the power of CNNs to extract rich and hierarchical features from input images. CNNs excel at learning complex patterns, allowing us to capture intricate details and essential visual cues from the data.

Our research integrates the U-Net architecture, a proven model in semantic segmentation tasks. U-Net's unique encoder–decoder structure enables precise semantic context analysis. By accurately capturing contextual information, we enhance the understanding of the relationships between different elements within the images.

A key innovation in our approach is the incorporation of a multi-head attention mechanism. This component enhances the model's focus on relevant image regions, allowing for adaptive feature weighting and more nuanced analysis. By employing attention mechanisms, our model can dynamically adjust its focus, improving both accuracy and efficiency.

The overall structure of the model is shown in Fig. 1.

CNN Layers

In our study, we utilized CNNs as a crucial component for processing robot images. This network employs convolutional operations, nonlinear activation functions, and pooling operations to gradually extract and learn features from the images. Here is the detailed explanation of each step with corresponding mathematical formulations:

(1) Convolution operation: convolution is a fundamental step in CNNs, involving the sliding of convolutional kernels over the input images. The mathematical expression for convolution at a specific position $\left(i, j\right)$ is defined as follows:

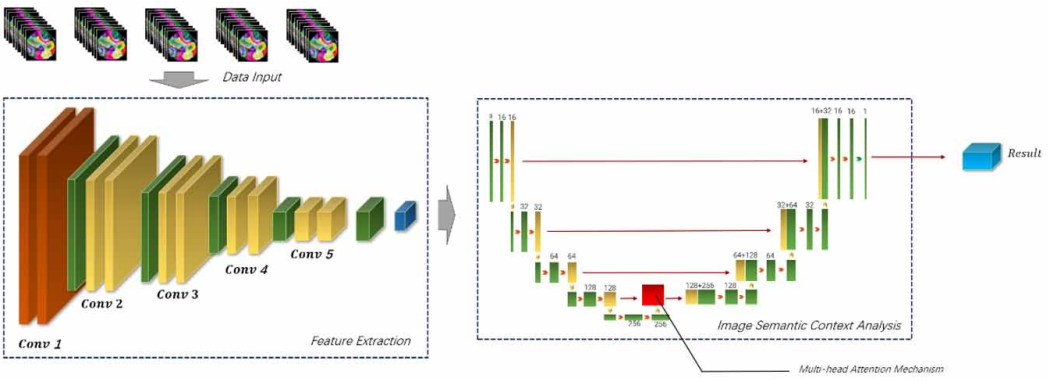
$$Y\left(i,j\right) = \sum_{m=-a}^{a}\sum_{n=-b}^{b} X\left(i+m, j+n\right) \times W\left(m,n\right) \tag{1}$$

Here, $Y\left(i,j\right)$ represents the corresponding position in the output feature map, $X\left(i+m, j+n\right)$ denotes the local region of the input image, and $W\left(m,n\right)$ represents the weights of the convolutional kernel. This operation captures local features such as edges and textures, playing a crucial role in feature extraction.

(2) Nonlinear activation function: following convolution, a nonlinear activation function, typically ReLU (rectified linear unit), is applied to introduce nonlinearity into the network. The mathematical expression for ReLU is defined as:

$$f\left(x\right) = \max\left(0, x\right) \tag{2}$$

Figure 1. Overall structure of our model



This step enhances the network's expressive power, enabling it to learn more complex image patterns.

(3) Pooling operation: pooling operations, usually performed after convolution, reduce the spatial dimensions of feature maps. We employ max pooling, expressed mathematically as:

$$Y\left(i,j\right)=\max_{m,n}\left(X\left(i\times s+m,j\times s+n\right)\right) \tag{3}$$

Pooling operations preserve essential features by retaining the maximum values within local regions, thereby decreasing computational complexity. This step enables the network to handle input images of varying sizes.

Through these processes, our CNN layer gradually extracts abstract features from robot images, providing robust support for subsequent image semantic analysis.

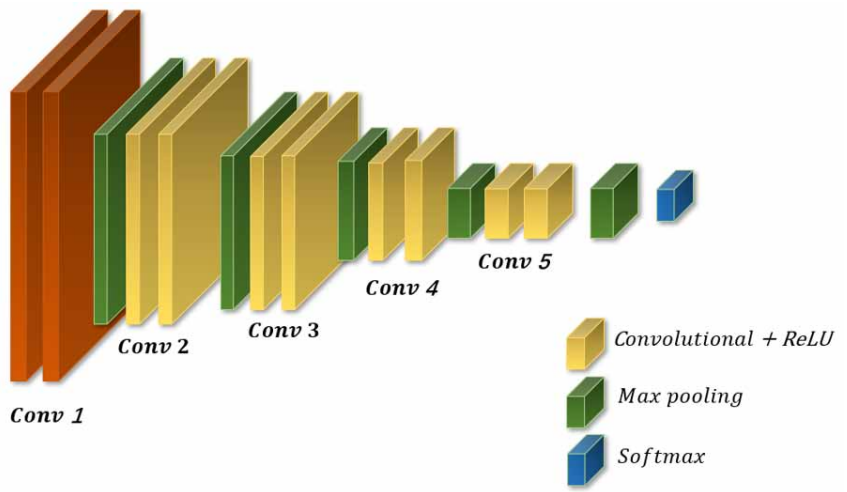
The architecture of the CNN layer is shown in Fig. 2.

As one of the innovations in this study, the CNN (Bao et al., 2021) demonstrates higher real-time efficiency compared to larger models such as transformers when dealing with small-scale image data. In tasks requiring rapid responses from robots, the CNN's swift feature extraction capability renders it a suitable choice.

U-Net Layers

Then we employed the U-Net model (Xue, 2021), specifically designed for image-segmentation tasks, characterized by its unique encoder–decoder architecture. The U-Net model's principle lies in its ability to efficiently handle image semantic segmentation through the encoder (downsampling path) and decoder (upsampling path) components. Here is a detailed explanation of the U-Net model's principles for processing robot image semantics, including both the encoder and decoder parts.

Figure 2. Structure of the CNN layer



(1) Encoder (downsampling path): the encoder is responsible for extracting high-level and low-level features from the input images. In the convolutional layers of the encoder, we utilized the convolution operation:

$$H_{i,j}^{l} = \sigma\left(W_l * H_{i,j}^{l-1} + b_l\right), \tag{3}$$

where $H_{i,j}^{l}$ represents the feature map of the l th layer, W_l denotes the convolutional kernel, b_l is the bias term, and σ represents the activation function. Subsequently, through pooling operations, the resolution of the feature maps is reduced, enabling the network to capture global image features effectively.

(2) Decoder (upsampling path): the decoder aims to restore the feature maps from the encoder to the original resolution. First, we utilized the deconvolution operation:

$$H_{i,j}^{l} = \mathrm{UpSampling}\left(H_{i,j}^{l-1}\right) \tag{4}$$

for upsampling the feature maps, where $\mathrm{UpSampling}\left(\cdot\right)$ denotes the operation commonly achieved through interpolation methods. Then, employing skip connections, the features from the encoder and decoder are concatenated, enriching the semantic information:

$$H_{i,j}^{l} = \mathrm{Concatenate}\left(H_{i,j}^{l-1}, H_{i,j}^{k}\right), \tag{5}$$

where $H_{i,j}^{k}$ represents the feature maps from the k th layer of the encoder.

(3) Loss function: to guide the network learning process, we utilized the cross-entropy loss function, which measures the discrepancy between the network outputs and the ground truth labels:

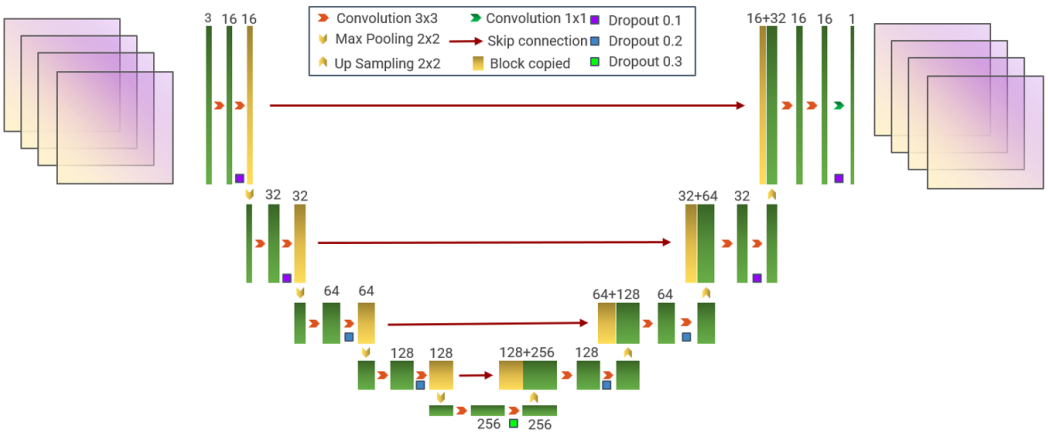
$$L = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}\log\left(\hat{y}_{i,c}\right), \tag{6}$$

where N is the number of samples, C is the number of classes, $y_{i,c}$ represents the true label's probability distribution, and $\hat{y}_{i,c}$ represents the model's output probability distribution.

Through these steps, the U-Net model can extract complex image features in the encoder and accurately reconstruct image details in the decoder, making it a powerful tool for processing semantic information in robot images. This architecture comprehensively integrates global and local features, enabling robots to accurately comprehend and segment intricate image semantic information. The architecture of the U-Net layer is shown in Fig. 3.

As one of the innovations in this study, the U-Net model's structure possesses a relatively small number of parameters, making it suitable for handling small sample data, a challenge commonly encountered in the field of robotics. Consequently, we opted to train the U-Net model on small-scale

Figure 3. Structure of the U-Net layer



datasets, thereby obtaining relatively accurate semantic analysis results. Additionally, the U-Net model is capable of processing high-resolution images, which proves highly beneficial for the high-definition image data acquired by robots in complex environments.

Multi-Head Attention Mechanism

Integrating a multi-head attention mechanism into the robot image semantic analysis algorithm based on U-Net enhances the network's ability to focus on different positions and feature channels, thereby improving the precision of semantic segmentation. Here are the steps, along with three equations, explaining the principle.

The multi-head attention mechanism enables the model to simultaneously focus on different parts of the input features, enhancing the network's representational capacity. In each attention head, linear transformations are applied to obtain queries (Q), keys (K), and values (V):

(1) Linear transformation: linear transformations are performed on the input features to obtain queries (Q), keys (K), and values (V):

$$Q = XW_Q, K = XW_K, V = XW_V \tag{7}$$

Here, X represents the input features, and W_Q, W_K, and W_V are learnable weight matrices.

(2) Attention computation: attention weights A are computed to represent the model's focus on different positions:

$$A = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) \tag{8}$$

Here, d_k denotes the dimensionality of queries or keys.

(3) Weighted summation: values are weighted by attention weights and summed to obtain the output Y of the multi-head attention:

$$Y = AV \tag{9}$$

(4) Integration of multi-head attention mechanism into U-Net: incorporate the multi-head attention mechanism after each deconvolution layer in U-Net's decoder part. Utilize the deconvolution output as input; perform linear transformations, attention computation, and weighted summation to obtain the multi-head attention output.

For instance, the multi-head attention output Y_l of the l th deconvolution layer in U-Net can be computed as follows:

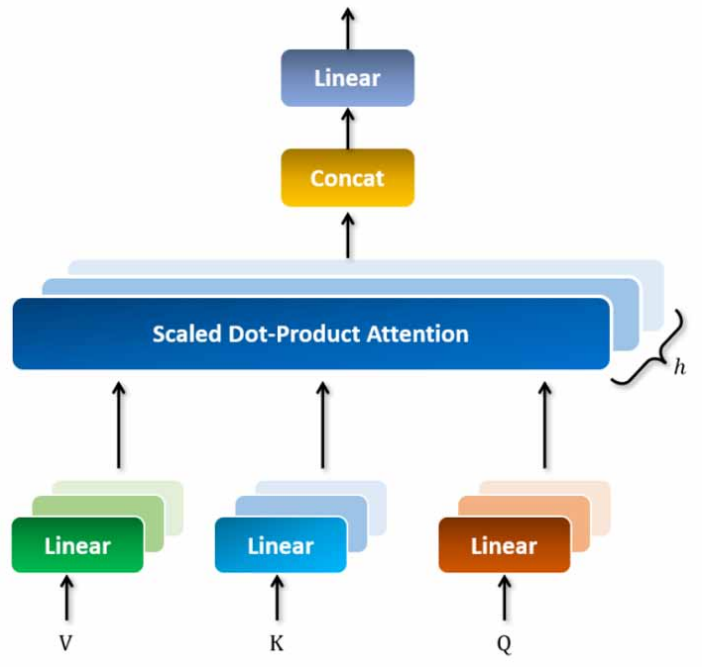
$$Y_l = \mathrm{MultiheadAttention}\left(X_l, X_l, X_l\right) \tag{10}$$

Here, X_l represents the output of the l th deconvolution layer, and $\mathrm{MultiheadAttention}\left(\cdot\right)$ denotes the computation process of the multi-head attention mechanism.

By adopting this approach, the multi-head attention mechanism is embedded into U-Net, enabling the model to handle different scales and semantic features of the image, thereby enhancing the performance of the robot image semantic analysis algorithm. The architecture of the multi-head attention mechanism is shown in Fig. 4.

As one of the innovations in this study, the application of the multi-head attention mechanism to address the task of robotic image semantic analysis has many advantages. Its primary advantage

Figure 4. Structure of the multi-head attention mechanism



lies in its inherent resistance to disturbances, allowing it to maintain focus on crucial information even in the presence of occlusion, uneven lighting, and similar conditions. This resistance enhances the robustness of the model, rendering it suitable for intricate real-world scenarios, which often pose challenges in robotic image processing tasks.

EXPERIMENT

Experimental Design

In this experiment, we aim to evaluate the performance of the proposed robot image semantic context analysis model, Rob-Att-UNet. First, we selected a publicly available robot image dataset containing diverse scenes and objects as the experimental foundation. Data preprocessing stages involved image standardization and data augmentation to enhance the model's robustness. We compared our model with 10 similar models sourced from the literature. The dataset was divided into training and validation sets, and model training and validation were conducted using cross-validation methods. Performance evaluation was based on metrics such as accuracy, recall rate, and AUC (area under the receiver operating characteristic curve), comprehensively assessing the model's classification accuracy and robustness. The analysis of the experimental results will provide crucial insights into the model's performance, strengths, and directions for improvement.

In terms of software environment, we utilized the Python programming language to implement machine-learning algorithms and employed libraries such as TensorFlow and PyTorch for model development and training. Data preprocessing and analysis were conducted using common data-processing libraries such as Pandas and NumPy. Additionally, for a more intuitive representation of experimental results, visualization tools including Matplotlib (Hunter, 2007) were utilized.

Regarding the hardware setup, the experiments were conducted on a high-performance computing cluster equipped with multiple graphics processing units to accelerate the training process of deep-learning models. The cluster environment possessed parallel-processing capabilities, enhancing the efficiency of model training. The data batch size is 4, and the training epoch is 100. The model parameter settings are shown in Tables 1 and 2.

Table 1. Parameter settings for the CNN and U-Net layers

Parameter	CNN	U-Net
Kernel Size	3x3, 5x5	3x3, 5x5
Stride	2	1
Padding	1	0
Activation Function	Sigmoid	ReLU
Pooling Size	3x3	3x3
Number of Conv. Layers	Variable	Multiple in encoder and decoder
Learning Rate	0.001	0.001

Table 2. Parameterization of the Multi-Head Attention Mechanism

Parameter	Multi-Head Attention
Number of Attention Heads	10
Attention Dimension	5

We evaluate the accuracy of the semantic context analysis of the robot images model using accuracy, recall, precision, F1-score, AUC, and training time(s):

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{11}$$

$$Recall = \frac{TP}{TP + FN} \tag{12}$$

$$Precision = \frac{TP}{TP + FP} \tag{13}$$

$$\mathrm{F1-score} = 2 \cdot \frac{Precision * Recall}{Precision + Recall} \tag{14}$$

The AUC metric is a measure used to assess the performance of classification models. In machine learning, when the performance of a binary classifier is being evaluated, a receiver operating characteristic (ROC) curve is typically plotted. The ROC curve is constructed with the true positive rate (also known as sensitivity or recall) on the vertical axis and the false positive rate on the horizontal axis. The AUC metric represents the area under the ROC curve.

Dataset

The five datasets in this article come from Robot Operating System (ROS) bags (Willis et al., 2022), KTH-IDOL (Luo et al., 2006), AI City Challenge (Huang, 2018), SUN RGB-D (McCormac et al., 2016), and ImageNet Robotics Vision Challenge (Lange, 2013). Table 3 describes the basic attributes of these five robotic image analysis datasets.

Table 3. Basic attributes of five robotic image analysis datasets

Dataset Name	Background and Purpose	Data Types	Key Features and Fields	Number of Samples
ROS Bags	Testing and development of robot perception and navigation algorithms	Camera images, LIDAR data, inertial sensor data	Sensor data, image data	1,000
KTH-IDOL	Research in object recognition, tracking, and scene understanding	RGB-D images, depth data, pose information, sensor data	Image data, depth information, pose information	1,000
AI City Challenge	Research tasks include vehicle detection, pedestrian recognition, and traffic analysis	Traffic camera images, vehicle and pedestrian bounding boxes, traffic flow data	Image data, bounding box information, flow data	1,000
SUN RGB-D	Research in scene understanding and object recognition	RGB-D images, object category labels, scene semantic segmentation	Image data, category labels, semantic segmentation information	1,000
ImageNet Robotics Vision Challenge	Research tasks include object detection, object recognition, and scene understanding	Images, object category labels, object bounding boxes	Image data, category labels, bounding-box information	1,000

ROS bags are a commonly used data-recording format within the Robot Operating System containing data from robot sensors such as camera images, LIDAR data, and inertial sensor data. These data are widely applied in testing and developing robot perception and navigation algorithms. ROS bags not only include data from static scenes but also capture robot movements, providing valuable experimental data for researchers to develop new algorithms enhancing robots' autonomy and perception capabilities in various environments.

The KTH-IDOL dataset is a multi-sensor dataset used for research in object recognition, tracking, and scene understanding. It includes RGB-D (red green blue depth) images, depth data, pose information, and other sensor data. These data are utilized in research related to robot visual perception both indoors and outdoors, serving as foundational data for tasks such as robot navigation, object manipulation, and environmental interaction.

The AI City Challenge dataset focuses on traffic analysis; it contains images of vehicles and pedestrians captured by traffic cameras, along with corresponding bounding boxes and flow data. These data are used for research tasks such as vehicle detection, pedestrian recognition, and traffic-flow analysis. Due to its real-world traffic-scene data, this dataset is used extensively in research related to intelligent traffic systems and autonomous vehicles.

The SUN RGB-D dataset comprises RGB-D images, object category labels, and semantic segmentation data, utilized for research in scene understanding and object recognition. These data reflect diverse indoor scenes, providing practical data support for robot applications in home and office environments.

The ImageNet Robotics Vision Challenge dataset provides a large-scale collection of images, including object category labels and bounding-box information. These data are used for robot-vision tasks such as object detection and object recognition. The dataset's diversity makes it a crucial resource for evaluating the performance of robot-vision algorithms.

These datasets offer rich experimental data for the development and performance evaluation of robot-vision algorithms, driving continuous advancements in robot technology in the fields of perception and cognition.

The Process of Model Parameter Tuning

Model parameter tuning involves the following five key processes:

(1) Thorough recording of each iteration's results: during the adjustment of model parameters, it is essential to meticulously document the outcomes of each training iteration, encompassing metrics such as loss function values and performance indicators. The real-time progress of the training can be monitored using the TensorBoard logging tool, and the results are stored in text files for subsequent analysis and comparison.
(2) Visual presentation and analysis: visualization tools such as Matplotlib and Seaborn are employed to graphically represent training metrics and loss function values. Through techniques like line plots and heatmaps, the performance trends of the model under various parameter combinations are analyzed. This visual representation provides an intuitive understanding of the model's behavior.
(3) Adjustment of configuration space: based on prior results and expertise, the search space for model parameters is adjusted. This adjustment involves defining reasonable ranges for parameters, including learning rates, batch sizes, and network structures. By targeted adjustment of parameter ranges, the efficiency and accuracy of parameter search are enhanced.
(4) Refinement of search algorithms: the grid search algorithm is utilized, involving an exhaustive exploration of all possible parameter combinations. The choice of an appropriate search algorithm depends on the specific characteristics of the problem and the availability of computational resources.
(5) Comprehensive derivation of final parameters and results: by synthesizing the outcomes of various experimental rounds, a holistic evaluation of model performance and training efficiency is conducted. Different parameter combinations are ranked based on predefined evaluation metrics

such as accuracy and loss function values. The model with the best performance is selected. Additionally, the model's generalization capability is considered to mitigate overfitting issues. The resulting parameter combination, obtained after meticulous tuning, is utilized for model deployment and applications.

Comparison Study Results and Analysis

Multimethod Comparison on a Single Dataset

We compare the Rob-Att-UNet model with the following 10 classical models, and for comparative analysis, all 11 models are run on the KTH-IDOL dataset.

(1) U-Net is a classic convolutional neural network designed for image segmentation tasks, featuring an encoder–decoder architecture, commonly employed in semantic image segmentation.
(2) DeepLab utilizes deep convolutional neural networks and dilated convolutions to capture image context effectively; often applied to image-segmentation tasks.
(3) Mask R-CNN combines object detection and semantic segmentation, allowing simultaneous object localization and precise segmentation mask generation.
(4) FCN is an end-to-end fully convolutional network capable of pixel-wise mapping from input images to semantic segmentation maps, frequently used for image segmentation tasks.
(5) SegNet is a lightweight convolutional neural network designed for real-time image segmentation, particularly suitable for tasks involving environment perception and navigation.
(6) PSPNet (Pyramid Scene Parsing Network) (Wang et al., 2021) utilizes a pyramid pooling structure to capture context information at different scales; applied in semantic image segmentation tasks.
(7) ENet (Paszke et al., 2016) is a lightweight convolutional neural network tailored for real-time semantic segmentation tasks, featuring an efficient network architecture.
(8) LinkNet (Chaurasia & Culurciello, 2017) is based on a fully convolutional network for image segmentation, enhancing segmentation accuracy through specific connectivity strategies.
(9) BiSeNet (Bilateral Segmentation Network) (Tsai & Tseng, 2023) integrates global and local information, employing bilateral convolutions to achieve image segmentation tasks effectively.
(10) HRNet (High-Resolution Network) (Sengupta & Srivastava, 2021) focuses on handling high-resolution images, preserving high-resolution features while performing multi-scale information fusion; suitable for image segmentation tasks.

The results of the comparative experiments are shown in Table 4 and Fig. 5.

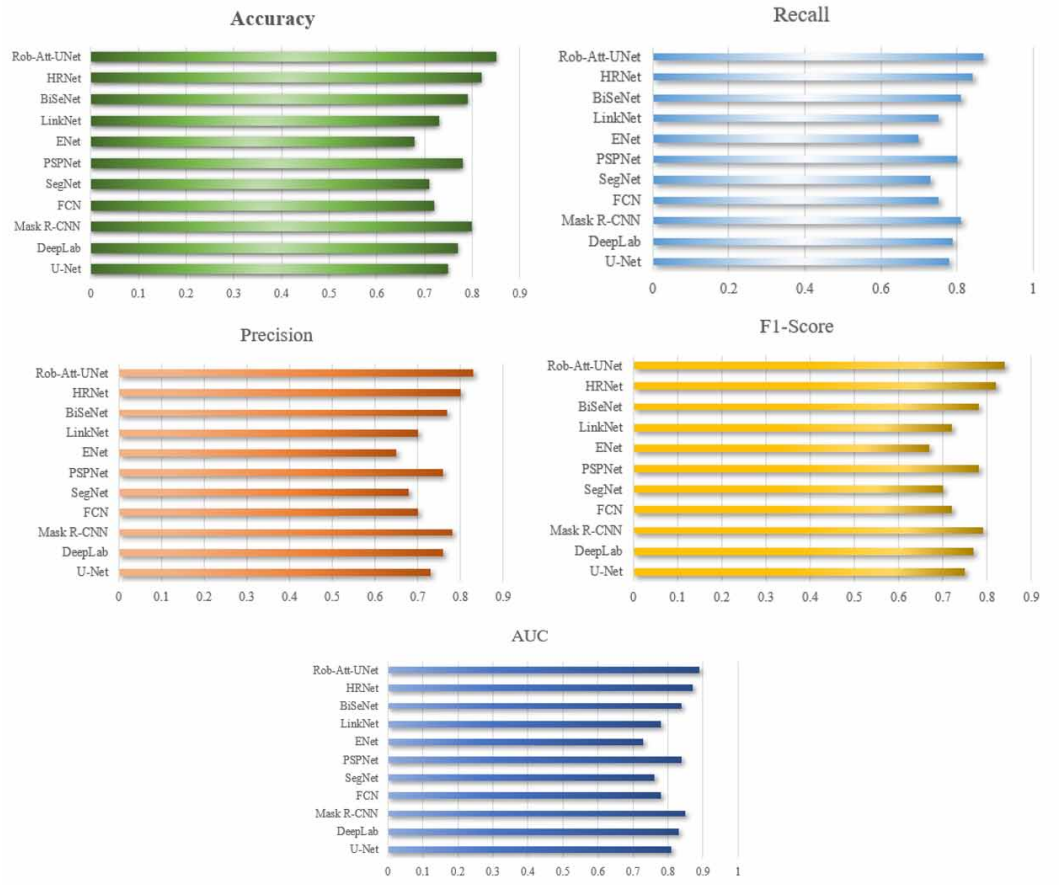
Based on the performance data table provided for the 11 models, we can observe variations in multiple metrics. In terms of accuracy, the Rob-Att-UNet model (0.85) excels, followed by Mask R-CNN (0.80), while the fundamental ENet model stands at 0.68. Under metrics like recall, precision, F1-score, and AUC, Rob-Att-UNet consistently maintains a relatively high level, showcasing its balanced performance across various metrics. Overall, Rob-Att-UNet demonstrates outstanding performance across multiple metrics, highlighting its robust capabilities in image semantic context analysis tasks. However, this also indicates that the model's performance is influenced not only by the algorithm itself but also by the characteristics of the dataset.

The Rob-Att-UNet model integrates various techniques and architectures, such as attention mechanisms, leveraging the advantages of different models to enhance overall performance. Additionally, the model has been optimized in terms of feature extraction and selection, enabling the CNN layers to capture essential information within the images more effectively, thereby enhancing the model's performance. Last, meticulous hyperparameter tuning has been conducted on the Rob-Att-UNet, ensuring the model achieves optimal performance for specific tasks.

Table 4. Results of the multimethod comparison experiment on a single dataset

	Accuracy	Recall	Precision	F1-Score	AUC
U-Net	0.75	0.78	0.73	0.75	0.81
DeepLab	0.77	0.79	0.76	0.77	0.83
Mask R-CNN	0.8	0.81	0.78	0.79	0.85
FCN	0.72	0.75	0.7	0.72	0.78
SegNet	0.71	0.73	0.68	0.7	0.76
PSPNet	0.78	0.8	0.76	0.78	0.84
ENet	0.68	0.7	0.65	0.67	0.73
LinkNet	0.73	0.75	0.7	0.72	0.78
BiSeNet	0.79	0.81	0.77	0.78	0.84
HRNet	0.82	0.84	0.8	0.82	0.87
Rob-Att-UNet	0.85	0.87	0.83	0.84	0.89

Figure 5. Results of the multimethod comparison experiment on a single dataset



However, the experimental result data still suggest some limitations of the model proposed in this paper. For example, the introduction of the multi-head attention mechanism increases the number of parameters in the model, leading to a certain degree of overfitting problem.

We also compare the training time of each method, as shown in Fig. 6.

From Fig. 6, it is evident that there are significant differences in the training times among different models. The Rob-Att-UNet model stands out with a training time of merely 17.6 seconds, whereas BiSeNet and HRNet require 22.8 seconds and 25.6 seconds, respectively.

These disparities in training time are likely attributed to the complexity of the model architectures and the number of parameters. Generally, models with more parameters and intricate structures demand additional time for training. However, it is essential to strike a balance between performance and training time when selecting a model. At times, relatively simpler models might complete training within relatively short periods. These models find utility in scenarios where real-time demands are high, especially under limited computational resources.

Comparison of Single Methods on Multiple Datasets

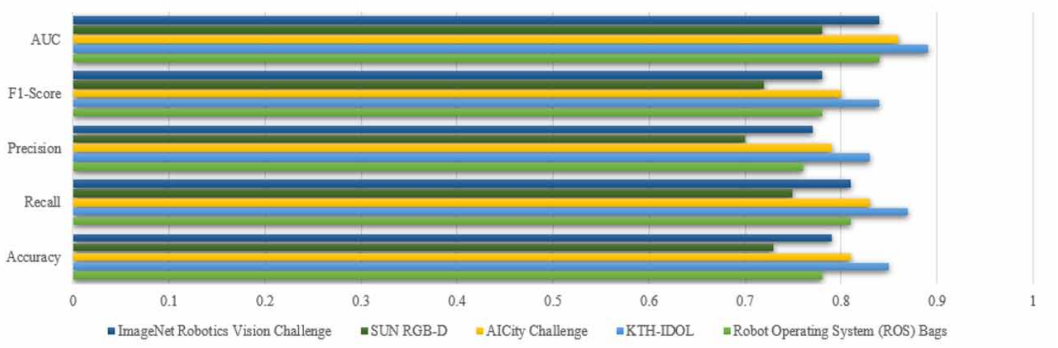
We compare the performance of the Rob-Att-UNet model in the five datasets mentioned above. The results of the comparative experiments are shown in Table 5 and Fig. 7.

Figure 6. The training time of each method

Table 5. Results of comparison experiment with single method on multiple datasets

	Accuracy	Recall	Precision	F1-Score	AUC
ROS bags	0.78	0.81	0.76	0.78	0.84
KTH-IDOL	0.85	0.87	0.83	0.84	0.89
AI City Challenge	0.81	0.83	0.79	0.8	0.86
SUN RGB-D	0.73	0.75	0.7	0.72	0.78
ImageNet Robotics Vision Challenge	0.79	0.81	0.77	0.78	0.84

Figure 7. Results of comparison experiment with single method on multiple datasets



In terms of the accuracy metric, the model excels on the AI City Challenge dataset with a score of 0.81, while performing at the lowest level on the SUN RGB-D dataset with a score of 0.73. Similar trends are observed in other metrics such as recall, precision, F1-score, and AUC, where the model exhibits outstanding performance on the AI City Challenge dataset, whereas its performance is relatively lower, around 0.76, on the KTH-IDOL dataset.

These disparities are likely attributed to the distinct characteristics and complexities of different datasets. The AI City Challenge dataset, focusing on the realm of intelligent transportation, possibly encompasses a plethora of rich and intricate traffic scenarios, thereby yielding superior model performance. In contrast, the SUN RGB-D dataset covers diverse indoor and outdoor settings, posing challenges for the model in handling such diversity and complexity, leading to comparatively lower performance.

Furthermore, it is evident that different datasets significantly impact model performance. A model excelling in a specific domain does not guarantee similar performance in other domains. Therefore, selecting datasets aligned with the task domain is crucial for accurate performance assessment and model selection. This analysis guides researchers in choosing appropriate datasets and models tailored to specific tasks, ensuring optimal performance in practical applications.

Ablation Study Results and Analysis

The ablation study compared the three models below.

(1) Original model: utilizing the traditional U-Net architecture without incorporating multi-head attention mechanism and CNN model. This model serves as our baseline for comparing with other experimental outcomes.
(2) CNN + U-Net model without attention: removing the multi-head attention mechanism from the U-Net structure while retaining the CNN model and keeping other conditions constant.
(3) Rob-Att-UNet: introducing the multi-head attention mechanism into the U-Net architecture to better capture crucial features within the images. This experiment aims to provide insights into the effectiveness of the multi-head attention mechanism in enhancing the model's performance and understanding its role within the U-Net framework.

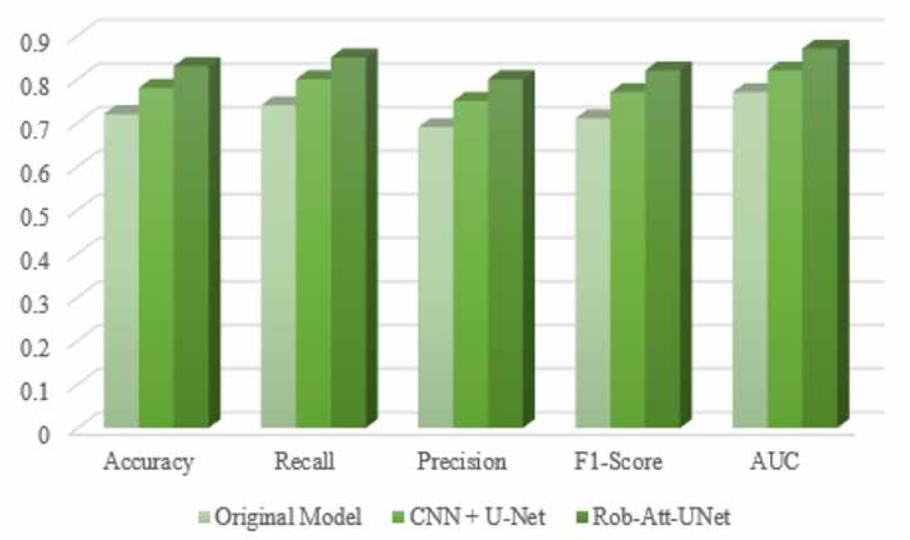
The results of the ablation study are shown in Table 6 and Fig. 8.

In terms of the five key metrics, namely accuracy, recall, precision, F1-score, and AUC, the Rob-Att-UNet model outperforms the other two models comprehensively. Specifically, Rob-Att-UNet achieves an accuracy score of 0.85, significantly surpassing both CNN + U-Net with 0.78 and the original model with 0.72. Furthermore, Rob-Att-UNet excels in recall, precision, F1-score, and

Table 6. Results of the ablation study

	Accuracy	Recall	Precision	F1-Score	AUC
Original Model	0.72	0.74	0.69	0.71	0.77
CNN + U-Net	0.78	0.8	0.75	0.77	0.82
Rob-Att-UNet	0.85	0.87	0.83	0.84	0.89

Figure 8. Results of the ablation study



AUC, with scores of 0.87, 0.83, 0.84, and 0.89, respectively, demonstrating a substantial lead over the other two models. These results emphasize the significant impact of the added CNN layer and multi-head attention mechanism in this study.

CONCLUSION

Our research endeavors to address the challenges faced by robots in semantic contextual analysis of images, including accurate comprehension of complex scenes and precise extraction of semantic information. To tackle these issues, we propose an innovative robot image semantic context analysis model that combines convolutional neural networks, the U-Net architecture, and the multi-head attention mechanism.

Initially, we incorporate CNNs as the foundational framework to extract fundamental features from the images. Subsequently, we employ the U-Net structure, allowing the network to learn richer semantic features, particularly excelling in handling image edges and detailed information. Concurrently, we introduce the multi-head attention mechanism, enabling the model to better focus on the interrelatedness among different regions within the image during semantic analysis, thereby enhancing the accuracy of semantic information and deepening contextual understanding.

In the selected models of the comparison experiment and ablation study, the Rob-Att-UNet model outperforms the others across all key metrics. It achieves an accuracy of 0.85, significantly surpassing other models. In other crucial metrics such as precision, F1-score, and AUC, Rob-Att-UNet also demonstrates outstanding performance, with scores of 0.83, 0.84, and 0.89, respectively, higher than

the other models. Furthermore, Rob-Att-UNet requires only a short training time of 17.6 seconds, compared to 22.8 seconds for BiSeNet and 25.6 seconds for HRNet, showcasing its faster training speed. These data highlight the dual superiority of the Rob-Att-UNet model in both performance and efficiency, making it an ideal choice for image semantic context analysis tasks.

By integrating three deep-learning techniques, our model has effectively captured the semantic information of complex scenes in images, providing robots with a more accurate and in-depth visual understanding. This research offers a viable approach for intelligent decision-making and interaction for robots across diverse environments, serving as a valuable reference for the advancement of robotics technology.

Overall, the innovativeness of this study lies in the judicious selection of a context-aware image semantic analysis technique that is both rational and efficient, thereby enhancing its applicability to the task of robot environmental image semantic analysis. According to our analysis, the primary challenges in robot image semantic analysis tasks stem from the real-time constraints imposed by downstream tasks such as robot navigation, placing strong limitations on the time available for image semantic analysis tasks. Additionally, the actual environments where robots operate are often highly complex, rendering single-headed attention mechanisms insufficient for the effective analysis of environmental information by robots. Therefore, this study employs a faster CNN model for image feature learning and integrates a multi-head attention mechanism to achieve a more comprehensive analysis of environmental image data.

With the widespread adoption of 5G technology and the development of related technologies such as edge computing, robot image semantic analysis technology is poised to become even more intelligent and efficient. In the future, the high-speed transmission and low-latency features of 5G will enable robots to rapidly acquire image data, while the implementation of edge computing will facilitate local image semantic analysis, reducing data transmission time and enhancing real-time performance and privacy security. The extensive impact of robot image semantic analysis technology is expected to continually expand, offering increasingly intelligent, efficient, and secure solutions across various fields.

Despite the notable performance of our proposed robot image semantic context analysis model in complex scenes, several challenges persist. The model's performance may be compromised when dealing with uneven lighting or shadow effects, as these factors can lead to information loss or distortion in the images. The model may also face accuracy issues when dealing with deformations and movements of nonrigid objects, which are crucial in real-world scenarios characterized by flexible and dynamic environments.

To overcome the aforementioned challenges, we plan to employ multiple strategies for enhancement. First, we intend to explore the integration of reinforcement learning (RL) techniques to enhance the model's adaptability under varying lighting and shadow conditions. By incorporating RL, the model can learn more robust feature representations through continual trial and error, thus improving its performance across diverse environmental settings. Second, we plan to introduce shape modeling and motion-estimation techniques to better handle deformations and movements of nonrigid objects. This will involve in-depth exploration of three-dimensional geometry and kinematic modeling, enabling the model to accurately capture shape and motion information of objects, thereby enhancing its understanding capabilities in real-world scenarios. Through these enhancements, we anticipate elevating our model to a new level, enabling it to excel in a broader and more intricate array of real-world contexts.

The environment in which robots operate encompasses not only visual information but also information from other modalities. The impact of environmental information from other modalities on tasks such as robot navigation is substantial. Therefore, there is a compelling need in the research on robot image semantic analysis to delve into the study of cross-modal environmental data learning and fusion. Future endeavors will be focused on exploring more effective ways to integrate information from diverse sensors, enabling robots to perceive their surroundings in a more comprehensive manner.

Additionally, research can be conducted to establish a deeper level of semantic understanding on the foundation of multimodal data fusion, enabling robots to interpret and respond to complex real-world scenarios more accurately.

The data available in the environment where robots operate is limited, and training an efficient robot image semantic analysis model requires a considerable amount of data. Future research will strive to explore transfer learning and few-shot learning methods in robot image semantic analysis. A critical unresolved challenge is to determine how a robot, when first applied in a new environment, can efficiently learn from small sample datasets using prior knowledge and experience. This entails the development of novel algorithms and models, allowing robots to extract universal semantic information from limited data and thereby adapt and execute tasks more rapidly when faced with new scenarios and tasks.

The environment in which robots operate differs significantly from controlled laboratory settings, often being more complex and unpredictable. Future research will aim to enhance the capability of robot image semantic analysis systems in context awareness and adaptive response to situations. The forthcoming work will emphasize the development of more intelligent and flexible robot systems, enabling them to better understand contextual information in complex environments and adapt autonomously to different scenarios. This includes real-time responses to environmental changes, dynamic recognition of scene features, and intelligent adjustments for diverse tasks.

The robot image semantic context analysis model proposed in this study not only addresses the challenges of semantic understanding in complex scenes but also provides crucial support for intelligent decision-making and interaction in practical applications of robotics technology. By integrating CNN, U-Net architecture, and the multi-head attention mechanism, we have successfully enhanced the performance of robot image semantic context analysis in complex scenes. Furthermore, this research offers valuable insights for future studies and applications. By improving the robot's visual perception and comprehension abilities, this study contributes to the advancement of intelligent robotics technology and the establishment of safer, smarter, and more efficient human–robot coexistence environments.

AUTHOR NOTE

The authors acknowledge research on the improvement of humanistic quality of special prosecutors (H7G170003) and on cultural Development Strategy of Shandong Nanxi Jinshi New Material Company (H7G210038).

ACKNOWLEDGMENT

The authors acknowledge research on the improvement of humanistic quality of special prosecutors (H7G170003) and on cultural Development Strategy of Shandong Nanxi Jinshi New Material Company (H7G210038).

REFERENCES

Balachandran, S., & Ranganathan, V. (2023). Semantic context-aware attention UNET for lung cancer segmentation and classification. *International Journal of Imaging Systems and Technology*, *33*(3), 822–836. doi:10.1002/ima.22837

Bao, T., Ren, N., Luo, R., Wang, B., Shen, G., & Guo, T. (2021). A BERT-based hybrid short text classification model incorporating CNN and attention-based BiGRU. [JOEUC]. *Journal of Organizational and End User Computing*, *33*(6), 1–21. doi:10.4018/JOEUC.294580

Cen, H. (2023). Target location detection of mobile robots based on R-FCN deep convolutional neural network. *International Journal of Systems Assurance Engineering and Management*, *14*(2), 728–737. doi:10.1007/s13198-021-01514-z

Chaurasia, A., & Culurciello, E. (2017). LinkNet: Exploiting encoder representations for efficient semantic segmentation. *arXiv*. /arXiv.1707.0371810.1109/VCIP.2017.8305148

Chen, M., & Zhang, L. (2023). The econometric analysis of voluntary environmental regulations and total factor productivity in agribusiness under digitization. *PLoS One*, *18*(9), e0291637. doi:10.1371/journal.pone.0291637 PMID:37708182

Cong, Y., Gu, C., Zhang, T., & Gao, Y. (2021). Underwater robot sensing technology: A survey. *Fundamental Research*, *1*(3), 337–345. doi:10.1016/j.fmre.2021.03.002

Dai, X., Zhang, Y., Jiang, J., & Bing, L. (2021). Image-guided robots for low dose rate (LDR) prostate brachytherapy: Perspectives on safety in design and use. *International Journal of Medical Robotics and Computer Assisted Surgery*, *17*(3), e2239. doi:10.1002/rcs.2239 PMID:33689202

Elmquist, A., Serban, R., & Negrut, D. (2022). Evaluating a GAN for enhancing camera simulation for robotics. *arXiv.* https://doi.org//arXiv.2209.0671010.48550

Gao, D., Zhu, J., Li, F., & Yang, Y. (2021). A hybrid and regenerative model chat robot based on LSTM and attention model. In H. Lu, S. Mu, & S. Nakashima (Eds.), *International Symposium on Artificial Intelligence and Robotics 2021* (Vol. 11884) (pp. 151–159). SPIE. doi:10.1117/12.2603769

Huang, T. (2018). Traffic speed estimation from surveillance video data: For the 2nd NVIDIA AI City Challenge track 1. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (pp. 161–1614). IEEE. doi:10.1109/CVPRW.2018.00029

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, *9*(3), 90–95. doi:10.1109/MCSE.2007.55

Kong, L. (2020). *Kinematic resolutions of redundant robot manipulators using integration-enhanced RNNs*. https://doi.org//arXiv.2008.0822810.48550

Kushwaha, V., Shukla, P., & Nandi, G. C. (2022). Generating quality grasp rectangle using Pix2Pix GAN for intelligent robot grasping. *arXiv*. https://doi.org//arXiv.2202.0982110.48550

Lange, U., Kampe, H., & Graeser, A. G. P. (2013). ImageNets—Framework for fast development of robust and high performance image processing algorithms. *Automatisierungstechnik*, *61*(3), 203–212. doi:10.1524/auto.2013.0019

Li, B., Ajjaji, O., Gigandet, R., & Nazir, T. (2023). The body image of social robots. *arXiv.* 10.1109/ARSO56563.2023.10187489

Li, Q., Jia, W., Sun, M., Hou, S., & Zheng, Y. (2021). A novel green apple segmentation algorithm based on ensemble U-Net under complex orchard environment. *Computers and Electronics in Agriculture*, *180*, 105900. doi:10.1016/j.compag.2020.105900

Lin, J., Li, Y., Xie, Y., Hu, J., & Min, J. (2022). Joint stiffness identification of industrial serial robots using 3D digital image correlation techniques. *Proceedings of the Institution of Mechanical Engineers. Part C, Journal of Mechanical Engineering Science*, *236*(1), 536–551. doi:10.1177/09544062211002878

Liu, Y., & Chen, M. (2023). The knowledge structure and development trend in artificial intelligence based on latent feature topic model. [early access]. *IEEE Transactions on Engineering Management*, ●●●, 1–12. doi:10.1109/TEM.2022.3232178

Lu, H., Liu, Q., Liu, X., & Zhang, Y. (2021). A survey of semantic construction and application of satellite remote sensing images and data. *Journal of Organizational and End User Computing*, *33*(6), 1–20. doi:10.4018/JOEUC.20211101.oa29

Luo, J., Pronobis, A., Caputo, B., & Jensfelt, P. (2006). *The KTH-IDOL2 database.* Technical Report CVAP304, KTH Royal Institute of Technology, CVAP/CAS. https://www.researchgate.net/publication/228386719_The_KTH-IDOL2_database

McCormac, J., Handa, A., Leutenegger, S., & Davison, A. J. (2016). SceneNet RGB-D: 5M photorealistic images of synthetic indoor trajectories with ground truth. *arXiv*. https://doi.org//arXiv.1612.0507910.48550

Nagata, F., Habib, M. K., & Watanabe, K. (2021). Transfer learning-based and originally-designed CNNs for robotic pick and place operation. *International Journal of Mechatronics and Automation*, *8*(3), 1. doi:10.1504/IJMA.2021.118430

Ning, E., Wang, C., Zhang, H., Ning, X., & Tiwari, P. (2023). Occluded person re-identification with deep learning: A survey and perspectives. *Expert Systems with Applications*, *239*, 122419. doi:10.1016/j.eswa.2023.122419

Okafuji, Y., Song, S., Baba, J., Yoshikawa, Y., & Ishiguro, H. (2022). Influence of collaborative customer service by service robots and clerks in bakery stores. *Frontiers in Robotics and AI*, *10*, 1125308. doi:10.3389/frobt.2023.1125308 PMID:37465719

Paszke, A., Chaurasia, A., Kim, S., & Culurciello, E. (2016). ENet: A deep neural network architecture for real-time semantic segmentation. *arXiv*. https://doi.org//arXiv.1606.0214710.48550

Patel, K., Bur, A. M., & Wang, G. (2021). Enhanced U-Net: A feature enhancement network for polyp segmentation. In *2021 18th Conference on Robots and Vision (CRV)* (pp. 181–188). IEEE. doi:10.1109/CRV52889.2021.00032

Sengupta, K., & Srivastava, P. R. (2021). HRNET: AI on Edge for mask detection and social distancing. *arXiv*. https://doi.org//arXiv.2111.1520810.48550

Shah, D., Eysenbach, B., Kahn, G., Rhinehart, N., & Levine, S. (2021). Rapid exploration for open-world navigation with latent goal models. *arXiv*. https://doi.org//arXiv.2104.0585910.48550

Singh, K. J., Kapoor, D. S., Thakur, K., Sharma, A., & Gao, X. Z. (2022). Computer-vision based object detection and recognition for service robot in indoor environment. *Computers, Materials & Continua*, *72*(1), 197–213. doi:10.32604/cmc.2022.022989

Sun, D., Zhao, H., Song, T., Liu, A., Cheng, J., Lio, Z., & Zhao, X. (2021). Learning hierarchical face representation to enhance HCI among medical robots. *Future Generation Computer Systems*, *118*(12), 180–186. doi:10.1016/j.future.2020.11.007

Tsai, T. H., & Tseng, Y. W. (2023). BiSeNet V3: Bilateral segmentation network with coordinate attention for real-time semantic segmentation. *Neurocomputing*, *532*, 33–42. doi:10.1016/j.neucom.2023.02.025

Turgut, K., Dutagaci, H., & Rousseau, D. (2022). RoseSegNet: An attention-based deep learning architecture for organ segmentation of plants. *Biosystems Engineering*, *221*, 138–153. doi:10.1016/j.biosystemseng.2022.06.016

Wang, H., & Chen, W. (2021). Task scheduling for transport and pick robots in logistics: A comparative study on constructive heuristics. *Autonomous Intelligent Systems*, *1*(1), 17. doi:10.1007/s43684-021-00017-9

Wang, P., Chen, H., Ma, G., Li, R., & Wang, X. (2021). Deep learning scheme PSPNet for electrical impedance tomography. In H. Huang, D. Zonta, & Z. Su (Eds.), *Sensors and Smart Structures Technologies for Civil, Mechanical, and Aerospace Systems 2021, Proceedings* (Vol. 11591). SPIE. doi:10.1117/12.2582437

WillisA. R.BrinkK.DippleK. (2022). ROS georegistration: Aerial multi-spectral image simulator for the Robot Operating System. *arXiv*. https://arxiv.org/abs/2201.07863

Xue, W. (2021). Skin lesion segmentation method based on U-Net with multi-scale and multi-dimensional feature fusion. *Journal of Jilin University*, *59*(1), 123–127. http://xuebao.jlu.edu.cn/lxb/EN/Y2021/V59/I1/123

Ye, S., Yao, K., & Xue, J. (2023). Leveraging empowering leadership to improve employees' improvisational behavior: The role of promotion focus and willingness to take risks. *Psychological Reports*, *00332941231172707*. doi:10.1177/00332941231172707 PMID:37092876

Ye, S., & Zhao, T. (2023). Team knowledge management: How leaders' expertise recognition influences expertise utilization. *Management Decision*, *61*(1), 77–96. doi:10.1108/MD-09-2021-1166

Zhang, C., Liu, G., Zhan, X., Shi, H., Cai, H., & Li, Y. (2021). Multiple object tracking algorithm based on Mask R-CNN. [Science Edition]. *Journal of Jilin University*, *59*(3), 609–618.

ZhaoS.OtaK.DongM. (2022). UAV base station trajectory optimization based on reinforcement learning in post-disaster search and rescue operations. *arXiv.* https://arxiv.org/abs/2202.10338

Zhenzhen, S., Zhou, Z., Wang, W., Gao, F., Fu, L., Li, R., & Cui, Y. (2021). Canopy segmentation and wire reconstruction for kiwifruit robotic harvesting. *Computers and Electronics in Agriculture*, *181*, 105933. doi:10.1016/j.compag.2020.105933