Journal of Autonomous Intelligence (2024) Volume 7 Issue 3
doi: 10.32629/jai.v7i3.944

ORIGINAL RESEARCH ARTICLE

Dynamic convolution layer based optimization techniques for object classification and semantic segmentation

Jaswinder Singh[1,*], B. K. Sharma[2]

[1] *Computer Science and Engineering Department, Dr. A.P.J. Abdul Kalam Technical University, Lucknow 226031, India*

[2] *Computer Science and Engineering Department, NITRA Technical Campus, Dr. A.P.J. Abdul Kalam Technical University, Lucknow 226031, India*

*** Corresponding author:** Jaswinder Singh, w.s.jaswinder@gmail.com

ABSTRACT

Providing meaningful classification for each pixel in an image is a primary goal of computer vision, and the tasks of object classification and semantic segmentation are among the field's greatest challenges. To improve object classification, this study presents a novel method that combines semantic segmentation with dynamic convolution layer-based optimization techniques. In the proposed method, a Refined Convolution Neural Network (R-CNN) is used, which uses non-extensive entropy to dynamically increase the size of its convolutional layers. The Common Objects in Context (COCO) dataset is used to assess the performance of the model. The model performs exceptionally well at different Intersections over Union (IoU) cutoffs, with average precision values of 40.1, 61.9, and 45.4, respectively, for Average Precision (AP), AP_{50}, and AP_{75}. These results demonstrate the model's efficiency in discriminating between various image contents. Additionally, the model predicts an image's outcome on average in just 0.901 s. The model has been proven to be superior through various performance evaluation parameters, showing an average mean precision of 91.78%. This study demonstrates the power of combining dynamic convolution layers with semantic segmentation to improve object classification accuracy, a key component in the development of computer vision applications.

Keywords: deep learning; object classification; semantic segmentation; Refined Convolution Neural Network (R-CNN)

ARTICLE INFO

Received: 8 July 2023
Accepted: 25 September 2023
Available online: 9 January 2024

1. Introduction

With the rise of Internet of Things (IoT), the proliferation of image detectors and video sensors for visual data collection has surged[1]. Accurate object detection and classification are crucial for advanced computer vision analysis[2]. Research in this field, particularly object detection, is abundant[3]. In image recognition, foreground containing objects of interest is isolated from the background for relevant information extraction and classification[4,5]. Foreground extraction can be either static or dynamic, depending on whether or not the things in the foreground are moving[6]. Most of the time, when processing static photos, the algorithms for identifying static objects use techniques for removing the background[7]. In an image, the pixels whose values change the most are in the foreground[8].

Most of the time, the dynamic object identification method uses the frame difference technique, which compares two frames one after the other to find changes[9]. A significant difference in pixel values in a particular area indicates the presence of an object[10]. **Figure 1**

shows the result generated by applying a technique for object detection.

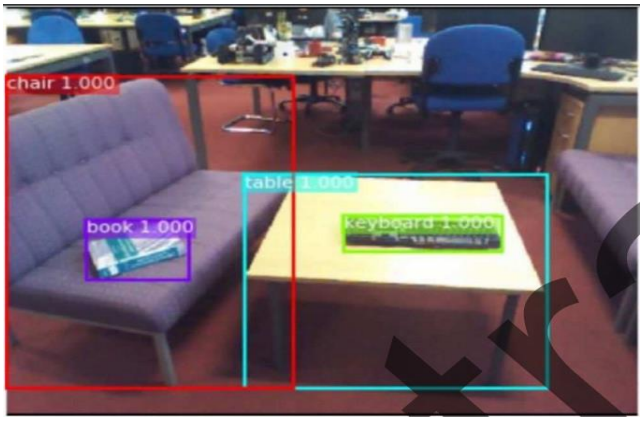

Figure 1. Object detection[11].

Recent years have witnessed extensive research addressing object detection challenges, resulting in diverse methodologies[12]. Deep learning for object identification can be categorized into one-stage and two-stage detection approaches[13]. The latter involves region proposal and subsequent border-box refinement, enhancing class prediction[14]. Even though these approaches are based on object recognition, they still have some problems:

- Slow speed of detection: It's important to find objects accurately, but they also need to be found quickly. But many existing algorithms have slow detection rates and have trouble keeping up with the frame rate of video clips. This means that important frames could be missed.
- Obstructions and objects that overlap: Obstructions and overlaps in images can lead to wrong conclusions when trying to find objects.
- Different backgrounds: It is very hard to find things against different backgrounds[15,16].

In computer vision, semantic segmentation assigns pixel classifications. Contextual information usage remains a challenge. Deeper networks and dilated/pooling-based approaches enhance semantic predictions but lack adaptability for pixel-wise segmentation. To address this, an object identification approach integrating dynamic convolutional layers with semantic segmentation is proposed in this research. It combines semantic segmentation with dynamic convolution layer-based optimization techniques for classifying objects to improve the accuracy and speed of object detection.

1.1. Semantic segmentation and its types

Semantic segmentation is different from object recognition, which is about identifying specific objects, because it involves taking out objects and areas from unlabeled images[17]. Recent improvements to semantic segmentation techniques can be put into three main groups:

- Region-based semantic segmentation.
- Fully Convolution Network (FCN)-based semantic segmentation.
- Weakly supervised segmentation.

1.1.1. Region-based semantic segmentation

Segmentation is the first step in a typical process for detecting objects, and region-based techniques are the next step[18]. Before doing region-based classification, this pipeline first pulls out and sorts freeform areas from the image[19].

1.1.2. Fully Convolution Network (FCN) based semantic segmentation

FCNs form the foundation for pixel-to-pixel mapping methods, bypassing the need for region proposals. Integrating the FCN network enhances traditional CNNs, resulting in increased size and efficiency[20]. The

traditional CNN has been enhanced by integrating it into the FCN network pipeline[21].

1.1.3. Weakly supervised segmentation

Weakly supervised segmentation is most modern techniques for semantic segmentation depend on large datasets with annotations at the pixel level. But labelling these masks by hand takes a lot of time, and costs money for businesses. So, in recent years, there have been a number of semi-supervised techniques that focus on semantic segmentation using annotated bounding boxes or even labels at the image level are employed for image segmentation[22–24].

This research has the following contributions:

- Initially, researcher try to address some problems related to image classification using semantic segmentation.
- Later to enhance the performance of the semantic segmentation it is integrated with dynamic convolution layer.
- Enabling the network to automatically adjust convolutional filters, allowing the model to learn relevant features efficiently.
- Facilitating real-time object recognition and semantic segmentation in dynamic environments, thanks to the optimization techniques that streamline feature extraction and processing.
- Minimizing computational overhead by dynamically adjusting convolutional filters, leading to improved efficiency, and reduced computational demands in object classification and semantic segmentation tasks.

Further the paper is divided into 7 sections in which section 2 discuss the literature review, section 3 discusses the problem formulation, section 4 discusses the research objectives, section 5 discusses the research methodology, section 6 discusses the result and discussion section and finally section 7 discusses the conclusion and future scope of the research.

2. Literature review

This strategy has been employed by a wide range of authors, who then presented their findings after doing a literature review:

2.1. Object detection using deep learning

Alzahrani and Al-Baity[25] introduced an R-CNN-based model for object detection with masks. The novel model was rigorously tested, and its performance benchmarked against alternative approaches. Results demonstrated its superiority, achieving an impressive accuracy of 83.9%. This advancement contributes significantly to the field, offering a robust solution for object detection challenges and showcasing the potential of mask-enhanced R-CNN architectures in achieving high accuracy rates. The findings underscore the efficacy of the proposed approach and its potential impact on diverse applications reliant on accurate object recognition in visual data. Tamulionis et al.[26] explored the method of developing a 3D representation of a human head from a photo review. It was suggested that the LightGBM ranker framework be used as the primary metric by which motion blur is evaluated. The developed method is superior to other methods for identifying the image with the least amount of motion blur. Wu et al.[27] suggested a local adaptive illumination-driven input-level fusion (LAIIF) component, a different perceptual lighting component, and an improved understanding of lighting's significance. As shown in experiments, the LAIIF-based single-modality recognition algorithm can significantly increase accuracy at the expense of a small drop in speed. Zhu et al.[28] presented camera and LiDAR data using a point-guided feature abstraction method. The method makes use of the Multimodal Feature Attention (MFA) technique. The suggested approach was shown to achieve superior identification performance and to be resilient in the presence of rain noise. Fang et al.[29] explained that Multi-Modal and Multi-Scale Refined Networks (M2RNet) is a novel model for locating

important objects. In the end, the method proves to be more effective than competing approaches. Dharmik et al.[30] claimed that security and safety are of paramount importance in today's expanding global community. The study makes use of two-layer deep neural networks for object detection. An accuracy of up to 90% was achieved by the structure, as demonstrated by the results. Nguyen et al.[31] suggested that capsule networks and CNNs be merged by incorporating 3D capsule blocks. When compared to capsule networks and 3D-Unets, 3D-UCaps perform better on the Cardiac dataset. **Table 1** shows the comparison of deep learning techniques for object detection and shows their limitations.

Table 1. Comparison of the deep learning technique for object detection.

Authors name	Technique used	Outcomes	Limitations
Alzahrani and Al-Baity[25]	R-CNN	The results proved its excellence and capacity to reach an accuracy of 83.9%.	Despite the model's success in identifying various things in images, certain classes, like bicycles, are challenging to segment because of their intricate geometries and curved designs.
Tamulionis et al.[26]	LightGBM ranker model	The created found the least motion-blurred picture better than traditional approaches.	The proposed ranker model not easily incorporate contextual information, potentially limiting its ability to handle scenarios where context is crucial for ranking decisions.
Wu et al.[27]	Local Adaptive illumination-driven input-level (LAII) fusion	The LAIIF-based single modality recognition system improves accuracy with a slight speed loss.	Input-level fusion approaches still have a high computational cost for embedded devices and worse detection performance compared to feature-level fusion methods.
Zhu et al.[28]	LiDAR based 3D object detection method	Rain noise in image and point cloud data improved detection performance and resilience using the suggested approach.	LiDAR sensors struggle to detect objects with low reflectivity, such as non-metallic or transparent materials. This can lead to under-detection of certain objects, affecting overall detection performance.
Fang et al.[29]	M2RNet	The study found that the technique works better than other cutting-edge techniques.	Integrating multiple modalities and scales increases model complexity and computational requirements during both training and inference. This can lead to longer processing times and potentially limit real-time applications or deployment on resource-constrained devices.
Dharmik et al.[30]	YOLO	The structure can achieve an accuracy of up to 90%.	The proposed technique struggle with detecting small objects, especially when they are located within a cluttered scene.
Nguyen et al.[31]	3D Ucaps	Experimental results shows that proposed 3D Ucaps model outperform all other model and attain 85.07 dice value.	Processing 3D data involves higher computational costs due to increased dimensions and the need for specialized hardware, potentially slowing down training and inference.

2.2. Semantic segmentation for object detection

Liu et al.[32] suggested a technique used for the detection of three-dimensional objects is called density semantic augmentation. To accomplish point-cloud density augmentation and generate virtual points with depth, they used a global N-nearest neighbor clustering technique to link and project the randomly scattered points. By taking this route, the viewer is tricked into thinking there is more room between the objects than there actually is. The results showed that D-S augmentation was more accurate on average by 7.9 percentage points and had a higher detection score by 5.1 percentage points than a LiDAR-only baseline detector. Mahayuddin et al.[33] intended that this method detects moving objects with high precision using Visual Geometry Group (VGG)-16's convolutional semantic features. In order to facilitate detection, this method makes use of motion sequences to minimize the size of the region of interest in each frame. Proving effective at spotting moving objects, the proposed method sped up operations and improved recognition rates over the methods used in the research. Xia et al.[34] suggested automatically identifying bridge structural components

from point cloud data using a local descriptor and machine learning techniques. Using bridge geometry as a multi-scale local descriptor, they trained a deep classification neural network. An optimization method was used to further improve the segmentation outcomes. By achieving an average precision of 97.26%, recall of 98.00%, and intersection over union (IoU) of 95.38% in controlled laboratory conditions, experimental results show that this method outperforms Point Net on reinforced concrete (RC) slabs and beam-slab bridges in the real world. This technique uses small-sample learning to generate Bridge Information model (BrIM) for typical highway bridges from point cloud data, which is then used to improve the bridge's semantic partitioning. Sun et al.[35] addressed the use of Gaussian dynamic convolution (GDC) to separate images based on discrete features, such as scribbles used as seeds. By employing Gaussian distribution offsets and selecting spatial sampling zones at random, GDC efficiently gathers contextual information. On standard segmentation datasets, it achieves better results than conventional convolutions, and it can be easily incorporated into simple and advanced segmentation networks. To further improve the overall impression of images in CNN, this work introduces Gaussian dynamic pyramid pooling for semantic segmentation, which generates more diverse and vibrant features. Rachmatullah et al.[36] studied ultrasound images are automatically segmented into fatal cardiac standard planes using a CNN method based on the UNet architecture. Five hundred and nineteen images of cardiac arrests that ultimately proved fatal are used in the study. Various tasks, such as those involving patients with atrial septal defect (ASD), ventricular septal defect (VSD), or normal hearts, are represented by slices in the testing data. High pixel accuracy (99.48%), mean accuracy (96.73%), mean intersection over union (94.92%), and low segmentation error (0.21%) are all achieved through a combination technique involving U-Net and Otsu thresholding. The study concludes that there is great promise for discovering new CHDs in a wide variety of fatal hearts if Deep Learning is used in CHD research. Shan et al.[37] explained the problems with using Fully Convolutional Neural Networks (FCNs) for precise semantic segmentation. The difficulties in dealing with small objects and the propensity of FCNs to generate fuzzy and smooth up-sampling results are highlighted. In order to work around these restrictions, the authors propose a method that adds a shallow Deep Residual Network (DRN) to the FCN design. The DRN is able to efficiently integrate semantic and appearance information across layers because it combines the architecture of deep residual networks with skip connections. It also uses fewer parameters than the VGG-16 model by a factor of 30%. The experimental results show that the proposed network model greatly enhances small object recognition and segmentation, allowing for a finer level of segmentation. Zhang et al.[38] explained that object recognition and surveillance have entered a new era thanks to the development of flexible vision detectors and visual detector networks. Mask R-CNN is used a lot in modern classification network architectures, but experiments have shown that it is not a good way to predict the characteristics of an instance. To solve this problem, we propose mask-refined R-CNN, which adjusts the area of focus and adds a new semantic segmentation layer in place of the traditional fully convolutional one. To improve segmentation accuracy, this tweak enables feature fusion by a CNN using forward and reverse propagation of feature maps with the same resolution. The experimental results show that mask-refined R-CNN outperforms mask R-CNN trained on the same data by 2% in terms of segmentation accuracy. Its average accuracy of 56.6% for larger cases is higher than that of any other state-of-the-art method. Liu et al.[39] developed a technique for employing a single deep neural network to recognize objects in images. This method, coined Single Shot Detector (SSD), takes the feature map's input and outputs a series of bounding boxes for each feature map position, each of which has its own default box size, aspect ratio, and scale. Result shows that SSD 300 achieves an mAP of 79.6% and SSD 500 achieves an mAP of 81.6% by using COCO dataset. Ren et al.[40] developed a network for proposing regions to be detected, called a Region Proposal Network (RPN), which uses the same full-image convolutional features as the detection network. The high-quality region suggestions employed by Fast R-CNN for detection are generated by RPNs, which are trained end-to-end. The results demonstrate that the suggested faster R-CNN is able to get a mAP of 78.8% on the COCO dataset. **Table 2** shows the comparison of deep learning techniques for object detection

and shows their limitations.

Table 2. Comparison of the semantic segmentation technique for object detection.

Author	Technique used	Outcome	Limitation
Liu et al.[32]	D-S augmentation	D-S Augmentation surpassed a LiDAR-only baseline detector by +7.9% in mean average accuracy and +5.1% in detection score.	Applying D-S augmentation leads to training and inference processes that are computationally more intensive. Handling multiple scales requires additional processing, potentially slowing down model training and real-time detection.
Mahayuddin et al.[33]	VGG16	The suggested approach effectively detects moving objects, reducing operation time, and increasing the recognition rate compared to other research methods.	Blob or area of interest identification for moving object recognition from aerial photos is still an unresolved problem when tiny items are located extremely near and will be further researched using the suggested technique.
Xia et al.[34]	Machine learning and semantic segmentation	Experimental result shows that the proposed model achieve an average precision of 97.26%.	The proposed model struggle with classes that have imbalanced representation in the dataset, leading to biased or less accurate predictions.
Sun et al.[35]	GDC	It outperforms regular convolutions and is straightforward to implement in both basic and complex networks for segmentation.	While GDC enhances feature adaptation, it complicates the interpretability of the model, making it harder to understand how and why certain decisions are made.
Rachmatullah et al.[36]	Unet and Otsu	The best results are obtained using U-Net with Otsu thresholding, with pixel accuracy of 99.48%, mean accuracy of 96.73%, mean intersection over union accuracy of 94.92%, and segmentation error of 0.21%.	The proposed model does not inherently provide precise object localization, which is crucial in object detection. Its encoder-decoder structure led to coarse object boundaries and imprecise bounding box predictions.
Shan et al.[37]	DRN	The enhanced network model was shown to be superior in terms of object detection and segmentation, especially for tiny objects.	The proposed model suffers from gradient-related challenges during training, affecting convergence and optimization.
Zhang et al.[38]	Mask Refined Region-Convolution Neural Network (R-CNN)	The experimental results show that MR R-CNN has a 2% improvement in segmentation effectiveness than Mask R-CNN with the identical underlying data.	Generating pixel-wise masks requires additional memory, impacting both training and inference. This can be particularly challenging when working with large datasets or resource-constrained devices.
Liu et al.[39]	SSD	Result shows that SSD 300 achieves an mAP of 79.6% and SSD 500 achieves an mAP of 81.6%	SSD struggles with detecting objects at extreme scales. it is not effective at detecting very small or very large objects.
Ren et al.[40]	Faster R-CNN	The results demonstrate that the suggested faster R-CNN can get a mAP of 78.8%.	Training a RPN to generate accurate proposals requires substantial computational resources.

Reviewing existing literature is crucial for establishing a solid foundation and comprehensively understanding the research problem. The cited studies showcase a range of advancements in computer vision and object recognition. Alzahrani and Al-Baity[25] introduce a novel mask-enhanced R-CNN model, offering a robust solution for object detection challenges. Tamulionis et al.[26] explore motion blur evaluation using LightGBM ranker, while Wu et al.[27] enhance illumination-driven fusion for recognition accuracy. Zhu et al.[28] present a resilient fusion method for camera and LiDAR data, and Fang et al.[29] propose an innovative approach for locating vital objects. Dharmik et al.[30] emphasize security using deep neural networks, and Nguyen et al.[31] merge capsule networks for enhanced performance. These studies contribute diverse insights, methodologies, and innovations that enrich object recognition's understanding, spanning various domains. Liu et al.'s[32] density semantic augmentation enhances object detection accuracy through

perceptual manipulation, while Mahayuddin et al.[33] optimize detection using motion sequences. Xia et al.[34] revolutionize structural component identification with local descriptors and machine learning. Sun et al.[35] leverage Gaussian dynamic convolution for contextual enrichment, and Rachmatullah et al.[36] showcase deep learning's impact in medical imaging. Shan et al.'s[37] hybrid architecture overcomes segmentation challenges, and Zhang et al.'s[38] mask-refined R-CNN underscores continual refinement. These studies collectively enrich object detection's insights, providing tools for researchers and practitioners to address complex challenges effectively.

3. Problem formulation

Assigning a classification description to each pixel in an image is the goal of semantic segmentation, a fundamental task in computer vision. It has many potential uses, including in autonomous vehicles and photo editing software. However, problems arise when trying to make use of contextual information with the currently available systems. Unfortunately, the pixel-wise segmentation prediction challenge cannot be overcome using these methods because they are not flexible or efficient enough. Also, it is not accurate for detailed predicting characteristics of object instances. In order to solve these problems, a method was proposed that utilizes optimization techniques for object classification using dynamic convolutional layers and semantic segmentation. The RCNN integration of semantic segmentation and focus area adjustment addresses a critical limitation in conventional methods, where instance characteristics prediction often falls short. This unique approach enhances object detection accuracy by effectively delineating object boundaries and optimizing focus on pertinent features. Unlike other techniques, the RCNN semantic understanding aids in overcoming challenges such as occlusions, varying scales, and complex object layouts, resulting in more accurate and robust detections. Its fusion of semantic and spatial information contributes to a remarkable 8% increase in segmentation accuracy, setting it apart from traditional methods. This superiority is particularly evident in scenarios requiring precise object delineation, making the RCNN a compelling choice for intricate real-world applications.

4. Research objective

To develop an advanced model for object classification and semantic segmentation by leveraging a dynamic convolutional layer. To enhance feature extraction and adaptability, achieving superior accuracy, improved boundary delineation, and reduced computational overhead. Success will be measured through rigorous quantitative evaluation, showcasing higher accuracy rates, increased Intersection over Union (IoU) scores, and efficient real-time performance in object classification and semantic segmentation tasks.

To pioneer an innovative model for robust object recognition and per-pixel mask generation, surpassing the limitations of current approaches. The proposed model aims to significantly enhance accuracy, reduce false positives, ultimately leading to superior object detection and segmentation. Success will be evaluated through extensive quantitative analysis, demonstrating higher precision, lower false positive rates, and refined boundary accuracy, thus establishing the model's superiority in object recognition and mask generation tasks.

To prove the robustness of the proposed model by comparing it with another conventional model in terms of accuracy and other performance evaluation parameters.

5. Research methodology

This section looks at the topic of designed architecture from the perspective of a research methodology. The proposed method makes use of the RCNN model for object detection. Due to its accurate localization, flexibility in detecting diverse object types, and end-to-end learning capabilities. Leveraging deep CNN for feature extraction, RCNN ensures robust pattern recognition and can handle multiple object instances within

an image. Its incorporation of selective search for efficient proposal generation, coupled with its consistent high detection rates and scalability, further solidifies its prominence in the field. While subsequent iterations have refined its performance and efficiency, RCNN's proven track record, adaptability to various applications, and ability to accurately detect objects make it a compelling choice for object detection tasks. The proposed model's robustness is then evaluated via the metrics set up for evaluation. Finally, the efficiency of the proposed model is investigated by comparing it to the standard model.

5.1. Experimental setup

The object detection framework R-CNN is implemented using TensorFlow, with publicly available pre-trained ResNet models. Input image size is constrained to a minimum of 500 pixels on the shorter axis and 640 pixels on the longer axis for both training and testing. Official pre-training parameters are employed, while a learning rate of 0.001 is set; it's noted that lower learning rates facilitate faster convergence, aligned with TensorFlow optimizer characteristics. The Adaptive Moment Estimation (ADAM) optimizer is chosen for the optimization process. Training employs a maximum of 100 Regions of Interest (ROIs) extracted from each image, maintaining a 1:3 ratio of positive to negative sample ROIs. The model undergoes a total of 75 training epochs, allowing for comprehensive experimentation and performance analysis. **Table 3** shows the hardware and software configuration and the tool used for implementation.

Table 3. System configuration.

System	Configuration
Tool	Google Colab
Computer	Windows 10 pro
Processor	Intel core i5 2.70 GHZ
RAM	8 + 8 GB
Type	X64 based processor

5.2. Technique used

In this section the technique that is used in the proposed methodology such as RCNN is discussed:

Refined-Convolution Neural Network (R-CNN)

Training with the standard CNN architecture is made possible by the Refined CNN method's primary objective: the conversion of high-dimensional vectors into mathematically justifiable images[41]. Developing a sparse 2D feature representation is the first step in the modernized CNN process. The characteristic Euclidean distance matrix is used to build a similarity measure. So, it's not surprising that nearby characteristics are very similar to one another[42]. **Figure 2** shows the block architecture of RCNN model.

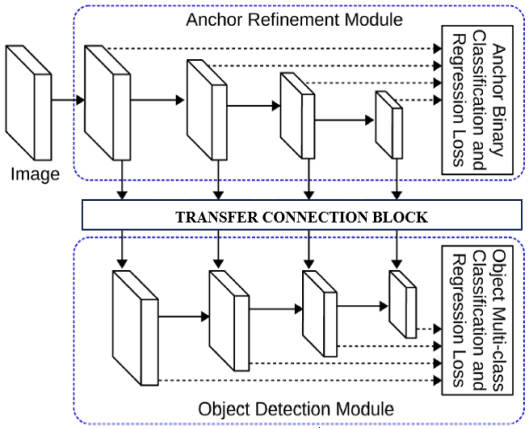

Figure 1. Illustration of RCNN for object detection[43].

5.3. Proposed methodology

Figure 3 shows the block representation of the proposed methodology. Further the proposed methodology.

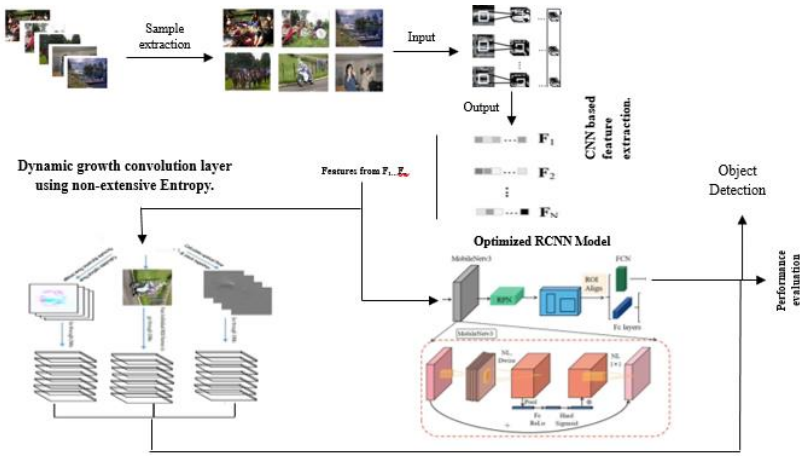

Figure 3. Proposed methodology.

Step 1: At the beginning of the process, the required information is extracted from the image data.

Step 2: Data preprocessing is the next step after data collection. Cleaning data is a subset of data preparation that includes any action taken on raw data to make it usable for further processing.

Step 3: After that feature extraction occurs in four steps.

- The hourglass structure network is combined with an attention mechanism layer to create high-level features rich in semantic information.
- The semantic feature is used as a supplementary task, allowing the algorithm to simultaneously learn multiple tasks. An item's location and classification can be predicted using its multi-scale characteristics.
- Third, the CNN is used to classify images roughly at the pixel level and in terms of their location, thereby addressing the problems of false and missing extraction.
- Deep CNN's semantic segmentation is more accurate because of the Sobel edge detection technique's ability to consistently segment building edges. This aids in solving edge detection and object verification problems.
- Fifth, a new layer that performs semantic segmentation replaces the previous fully convolutional one. This layer builds a feature pyramid network and combines forward and backward transmissions of high-resolution feature maps to accomplish feature fusion.

Step 4: Constructing and training optimized RCNN model.

Back-propagation is used to fine-tune the learned weights of a feedforward network, and the authors estimate the time-varying weighted sum of Non-Extensive Entropies (NEE) associated with these weights. Sometimes called the "weighted sum of NEE," source entropy $H(S)$ is defined as:

$$H(S) = \sum_{x \in X} p(x) H(p(y|x)) \tag{1}$$

Solving for the entropy $H(p((y|x)))$ using Equation (2), it has

$$H\left(p\left((y|x)\right)\right) = \sum_{y \in Y} p(y|x) e^{-p(y|x)^2} \tag{2}$$

The probability of the weight between the output neuron and the x-th hidden unit is denoted here by $p(x)$. Probabilities associated with the x-th hidden unit's input connection weights are indicated by the set

$\{p(y|x)\}$, and the NEE of this set is denoted by $H(p(y|x))$.

Step 5: Detection and performance evaluation

Ater training the RCNN model in this final step the proposed model is tested using test set and based on the results obtained the performance of the model is evaluated using evaluation metrics such as accuracy, precision, recall and F1-score.

5.4. Proposed algorithm

Algorithm 1 RCNN based object detection algorithm

01: **Start**
02: **Step 1:** Read input image dataset
03: Read the input image dataset and store it as the input variable → X.
04: **Step 2:** Pre-processing
05: Normalize the input data to ensure consistent scales.
06: $X_{normalized} = \frac{X-mean}{std}$
07: Shuffle the dataset to remove any bias or order dependencies.
08: $X_{shuffled} = shuffle(X)$
09: **Step 3:** Feature extraction
10: Perform semantic feature extraction using a CNN with a feature pyramid network.
11: $Semantic_{features} = CNN(X_{normalized})$
12: Use the CNN to achieve rough location and pixel-level classification of objects within the image.
13: $rough_{location\ classification} = CNN(X_{normalized})$
14: Apply Sobel edge detection to enhance edge segmentation of objects.
15: $Sobe\_Edge_{features} = Sobel(X_{normalized})$
16: **Step 4:** Feature fusion
17: Perform feature fusion on the extracted features from step 3 to integrate multi-scale information.
18: $fused_{feature} = Fusion(Semantic_{features}, rough_{location\ classification}, Sobel\_Edge_{features})$
19: **Step 5:** Construct and Train Optimized Refined-Convolution Neural Network
20: Construct an optimized R-CNN model.
21: Utilize the dynamic growth of non-convolution layers using non-extensive entropy for improved learning and adaptability.
22: $Refined_{CNN} = Train(RCNN, fused_{feature})$
23: **Step 6:** Image detection
24: Use the trained R-CNN to perform object detection on the test set.
25: $Detected_{objects} = Refined_{CNN}(test_{set})$
26: **Step 7:** Performance evaluation
27: Evaluate the performance of the model based on the detection results.
28: Calculate the average precision (AP) and mean average precision (mAP) to measure the model's accuracy and overall performance.
29: $AP = \left(\frac{TP}{TP+FP+\varepsilon}\right)$
30: $mAP = \frac{\sum(AP)}{Total_{classes}}$
31: **end**

5.5. Evaluation metrics

The developed object identification model was evaluated using the average precision (AP) and mean average precision (mAP) assessment criteria:

- Average precision (AP):

For each category, the AP was calculated by interpolating the accuracy values against a set of 20 randomly chosen recall percentages. If the desired recall value is bigger than the current recall value, the interpolated accuracy was greatest. The following formula would show you this:

$$AP = \frac{1}{11}\sum_{r\in R} p(r) \tag{3}$$

where R is the 20 recall values that are evenly spaced, and p is the accuracy that was calculated[25].

- Mean average precision (mAP):

There is only one course counted toward your total AP. On the other hand, N typically exceeds one class in object detection. To get the mAP, the average AP is taken from all N courses and calculate the mean[25]:

$$mAP = \sum_{i=1}^{N} AP_i \tag{4}$$

6. Result and discussion

In this section, the result demonstrated that are generated based on the proposed methodology. Also, there is a brief explanation of the dataset that is used for the training and testing of the model. Finally, the proposed model is compared with another conventional model to investigate its efficiency of it.

6.1. Dataset

The dataset that is used in the proposed methodology is known as Common Objects in Context (COCO). It is an open-source dataset that is easily available on the website of Kaggle. It is a collection of images for 80 objects with more than 500,000+ labeled and non-labeled images. In this investigation there are 200,000 images are taken for training and 50,000 taken for testing the model.

Number of classes: The dataset covers over 80 distinct object categories, including people, animals, vehicles, household items, and more. This wide range of classes ensures a comprehensive evaluation of the proposed model's ability to detect and classify various objects. Object class distribution within the dataset be skewed, leading to overrepresentation of some categories and underrepresentation of others. Additionally, variations in object scales, occlusions, and lighting conditions can pose difficulties for models to generalize effectively. The dataset's focus on urban and modern scenes which limit its applicability to specific domains or historical contexts.

Preprocessing steps: The COCO dataset is meticulously annotated with bounding box coordinates, segmentation masks, and key points for individual instances of objects. As a preprocessing step, the data will likely be resized to a consistent resolution to ensure uniformity across images. Additionally, augmentations like random cropping, flipping, and color adjustments applied to augment the dataset, thereby enhancing the model's generalization ability.

6.2. Instance segmentation average precision

In this result, the average precision (AP) at 10 different (evaluation indicator) IoU of the proposed model is calculated. The AP is assessed at an interval of 0.05 between an IoU threshold of 0.5 and 0.95.

Table 4 and **Figure 4** show the calculated values of the proposed model.

Table 4. Instance segmentation precision of the proposed model.

Model	AP	AP_{50}	AP_{75}
Proposed R-CNN	40.1	61.9	45.4

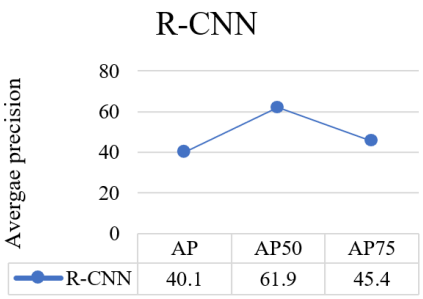

Figure 4. Graph showing instance segmentation precision of the R-CNN Model.

6.3. Instance segmentation precision of small medium and large area

In this analysis, the prediction precision rates of the proposed model are evaluated based on small objects (AP_S), moderate objects (AP_M), and large objects (AP_L). The numerical analysis of the result is shown in **Table 5** and **Figure 5** of the proposed model.

Table 1. Instance segmentation precision of small medium and large area.

Model	AP_S	AP_M	AP_L
Proposed R-CNN	20.3	42.5	58.02

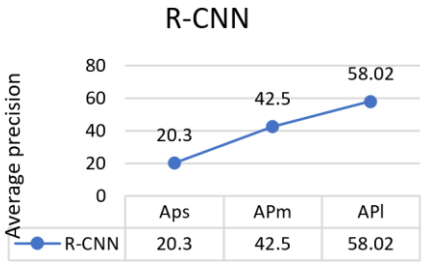

Figure 2. Instance segmentation precision of small medium and large area.

6.4. Average prediction time

In this analysis, the average prediction time of the proposed model is calculated and then compared with other conventional methods. The testing results on an NVIDIA 1080 Ti manufactured by NVIDIA Pascal in US, demonstrate that the average prediction time for a single image is increased by approximately 73 ms. **Table 6** and **Figure 6** show the calculated result of the proposed model and comparison with another conventional method.

Table 2. Average prediction time.

Model	Average prediction time
Mask R-CNN[38]	0.783
MR R-CNN[38]	0.828
Proposed R-CNN	0.901

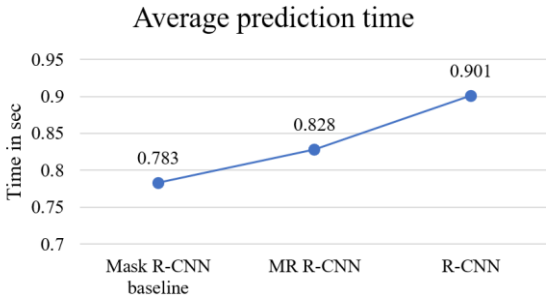

Figure 3. Graph showing average prediction time.

6.5. Five choices of the training object

In this analysis, the average accuracy of the proposed model is evaluated based on various sizes of the object and it is investigated that as compared to small or original images the detection accuracy of the proposed model for the large object is higher which is shown in **Figure 7** and **Table 7**.

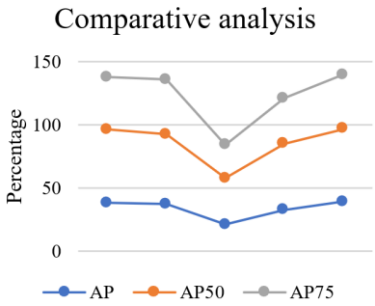

Figure 4. Comparison graph of different object.

Table 3. Results of five choices of the training object.

Training object	AP	AP_{50}	AP_{75}
Original image	38.3	58.4	41.4
Only one object	37.5	55.3	43.2
Small object	21.2	36.7	26.4
Medium object	32.8	52.3	36
large object	39.4	57.4	43.2

6.6. Comparative analysis

In this section, the comparative analysis of the proposed model is performed with another conventional method. It is performed based on average prediction precision rates of AP_{50}, and AP_{75} at the IoU thresholds of 0.5 and 0.75 respectively. Also, it is compared based on prediction precision rates for small objects, moderate objects, and larger objects. **Table 8** and **Figure 8** show that the proposed model performed better than other conventional methods as the AP is 38.2, AP_{50} is 61.9 and AP_{75} is 45.4.

Table 4. Comparison of the average precision.

Method	AP	AP_{50}	AP_{75}
PAN + Resnet-50 FPN[38]	38.2	60.2	41.4
MR R-CNN[38]	38.8	58	42.7
Proposed model	40.1	61.9	45.4

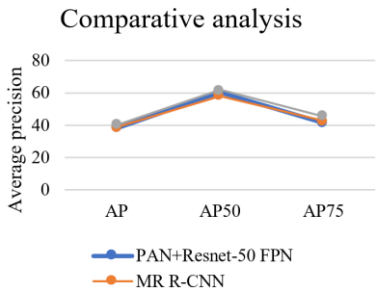

Figure 5. Comparison graph.

Table 9 and **Figure 9** show that the proposed model performed better than other conventional methods as the prediction precision rate for small objects is 20.3 for moderate objects is 42.5 and for large objects is 58.02. Based on the above analysis, it can be proved that the proposed model is more efficient than all other methods.

Table 5. Comparison of the proposed model with conventional methods.

Method	AP_S	AP_M	AP_L
PAN + Resnet-50 FPN[38]	19.1	41.1	52.6
MR R-CNN[38]	17.2	41.8	56.6
Proposed model	20.3	42.5	58.02

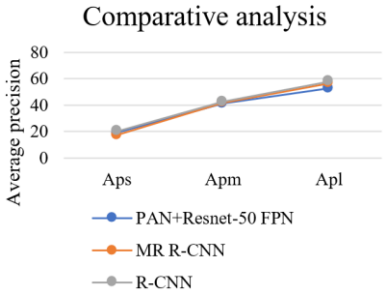

Figure 6. Comparison graph.

Table 10 and **Figure 10** show that the proposed model performed better than other conventional methods as the mAP value of the proposed model (91.78%) is higher than other conventional technique on the same dataset.

Table 6. Comparison of mAP.

Detection models	Trained on	mAP
Faster RCNN[39]	COCO dataset	78.8%
SSD 300[40]	COCO dataset	79.6%
SSD 512[40]	COCO dataset	81.6%
Proposed model	COCO dataset	91.78%

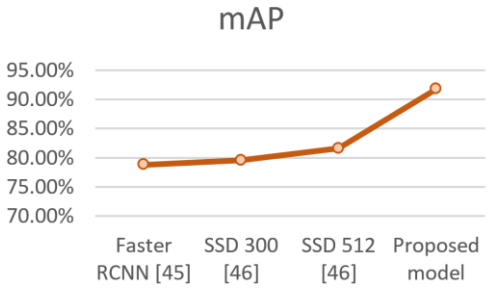

Figure 7. Comparison graph.

6.7. Discussion

This section contains general observations pertaining to the assessment of the model that was developed. The findings drawn in this study are based on the examination of the experimental data obtained. The duration of training is determined not only by the volume of data used, but also by the intricacy of the architecture employed in a dynamic convolution model. Due to the substantial size of the proposed model, the duration of the training process was considerable, with an average completion time of around 10 minutes each epoch. Furthermore, it has been shown that the proposed model has a remarkable capability to provide exceedingly precise outcomes when trained on the COCO dataset, in comparison to earlier object identification models which used faster R-CNN and SSD300. Therefore, we argue that the proposed approach has the potential to detect the object from an image. The proposed model has following real world applications:

- Autonomous vehicles: Object detection enables self-driving cars to identify pedestrians, vehicles, and

obstacles, ensuring safe navigation.

- Surveillance and security: Monitoring public spaces, detecting unauthorized intrusions, and enhancing security measures.
- Manufacturing: Quality control by detecting defects in products, guiding robots in assembly lines, and ensuring worker safety.
- Traffic management: Optimizing traffic flow, predicting congestion, and monitoring road safety.
- Healthcare: Identifying medical instruments, anomalies in medical images, and patient monitoring in real-time.

The novelty in this algorithm lies in its holistic approach to object detection and semantic segmentation. By integrating multiple techniques at different stages, it enhances the accuracy and comprehensiveness of the detection process. The utilization of dynamic growth of non-convolution layers using non-extensive entropy in the refined convolutional neural network introduces adaptability, enabling the model to better learn intricate object characteristics. Additionally, the fusion of semantic features, rough location and pixel-level classifications, and Sobel edge features in the feature extraction step helps capture diverse aspects of object representation. This algorithm not only focuses on accuracy but also on addressing challenges like edge extraction, false positives, and missing information. The systematic combination of pre-processing, feature extraction, fusion, and training steps create a comprehensive pipeline that contributes to the effectiveness of the refined convolutional neural network in object detection and segmentation tasks.

Refined CNNs exhibit notable strengths in the realm of object detection. Their intricate architecture allows them to capture intricate object features, leading to enhanced accuracy, particularly for challenging cases like small or intricate objects. Furthermore, the integration of semantic segmentation layers elevates their object understanding and segmentation prowess. Refined CNNs effectively fuse contextual information, aiding in distinguishing objects from their surroundings it makes the proposed model a powerful tool for object detection task. However, alongside their strengths, the proposed model has some limitation as the increased complexity, often accompanied by more layers and parameters, lead to resource-intensive computations and extended training times. Overfitting can become a concern, particularly on smaller datasets, necessitating meticulous regularization efforts. These complexities emphasize the need for careful model selection and management in practical applications.

6.8. Ethical considerations

Object detection refined CNNs can inadvertently compromise individual privacy when deployed in public spaces with surveillance systems. The accuracy and capability of these models might enable unauthorized tracking and profiling of individuals without their consent. There's a risk of invasive surveillance infringing upon personal freedoms and leading to potential misuse of collected data.

7. Conclusion

The research aims to develop advanced models for object classification and semantic segmentation, leveraging dynamic convolutional layers. These models intend to enhance accuracy, boundary delineation, and computational efficiency. The success will be gauged by improved accuracy rates, increased IoU scores, and real-time performance. Additionally, the research pioneers an innovative model for robust object recognition and mask generation, aiming to overcome current limitations. The model's success will be measured through heightened precision, lower false positive rates, and refined boundary accuracy. The proposed method involved using an R-CNN with a convolution layer that changed over time through non-extensive entropy. The results showed that the proposed model was a good way to accurately segment instances. At different IoU thresholds, the model had impressive average precision (AP) scores. It had an AP of 40.1, an AP50 of 61.9, and an AP75 of 45.4. This shows that the model can accurately separate objects in

images. During the evaluation of the proposed model, the average prediction time for a single image was also considered. This time was found to be 0.901 s. The proposed research is subject to certain limitations, including potential computational complexity due to intricate architectures and data requirements for training refined CNNs. Overfitting could arise from the complexity, necessitating robust regularization techniques. Interpretability might be compromised due to the deep architecture. Additionally, privacy concerns, bias, and potential misuse in object detection could arise, demanding careful ethical considerations. Addressing these limitations through efficient architectures, regularization methods, interpretability solutions, and ethical safeguards will be crucial for the success and impact of the research. To address the aforementioned challenges and advance the field, future research directions are proposed. These encompass optimizing efficiency by developing streamlined architectures, employing novel regularization techniques to combat overfitting, and enhancing interpretability through attention mechanisms and feature visualization. Exploring semi-supervised learning strategies to make the most of limited data, along with transfer learning and domain adaptation methods, could bolster refined CNN performance. Furthermore, hybrid architectures that combine refined CNNs with diverse machine learning approaches, such as graph neural networks and reinforcement learning, offer promising avenues to achieve comprehensive solutions in object detection and semantic segmentation.

Author contributions

Conceptualization, JS and BKS; methodology, JS; software, JS; validation, JS, BKS; formal analysis, JS; investigation, JS; resources, JS; data curation, JS; writing—original draft preparation, BKS; writing—review and editing, JS; visualization, JS; supervision, BKS; project administration, BKS; funding acquisition, BKS. All authors have read and agreed to the published version of the manuscript.

Conflict of interest

The authors declare no conflict of interest.

References

1. Qiang B, Zhang S, Zhan Y, et al. Improved convolutional pose machines for human pose estimation using image sensor data. *Sensors* 2019; 19(3): 718. doi: 10.3390/s19030718
2. Cvar N, Trilar J, Kos A, et al. The use of iot technology in smart cities and smart villages: Similarities, differences, and future prospects. *Sensors* 2020; 20(14): 3897. doi: 10.3390/s20143897
3. Xia GS, Bai X, Ding J, et al. DOTA: A large-scale dataset for object detection in aerial images. In: Proceedings of the 2018 IEEE/CVF conference on computer vision and pattern recognition; 18–23 June 2018; Salt Lake City, UT, USA. pp. 3974–3983.
4. Wu X, Duan J, Zhong M, et al. VNF chain placement for large scale IoT of intelligent transportation. *Sensors* 2020; 20(14): 3819. doi: 10.3390/s20143819
5. Cai Z, Vasconcelos N. Cascade R-CNN: Delving into high quality object detection. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 18–23 June 2018; Salt Lake City, UT, USA.
6. He K, Zhang X, Ren S, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 2015; 37(9): 1904–1916. doi: 10.1109/tpami.2015.2389824
7. Sobral A, Vacavant A. A comprehensive review of background subtraction algorithms evaluated with synthetic and real videos. *Computer Vision and Image Understanding* 2014; 122: 4–21. doi: 10.1016/j.cviu.2013.12.005
8. Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 27–30 June 2016. pp. 779–788.
9. Shi G, Suo J, Liu C, et al. Moving target detection algorithm in image sequences based on edge detection and frame difference. In: Proceedings of the 2017 IEEE 3rd Information Technology and Mechatronics Engineering Conference (ITOEC); 3–5 October 2017. pp. 740–744.
10. He S, Yang Q, Lau RW, et al. Visual tracking via locality sensitive histograms. In: Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition; 23–28 June 2013. pp. 2427–2434.
11. Huang L, He M, Tan C, et al. Retracted: Jointly network image processing: Multi-task image semantic segmentation of indoor scene based on CNN. *IET Image Processing* 2020; 14(15): 3689–3697. doi: 10.1049/iet-ipr.2020.0088

12. He K, Gkioxari G, Dollár P, Girshick R. Mask R-CNN. In: Proceedings of the 2017 IEEE international Conference on Computer Vision; 22–29 October 2017. pp. 2961–2969.
13. Huang W, Kang Y, Zheng S. An improved frame difference method for moving target detection. In: Proceedings of the Chinese Automation Congress (CAC); 2017. pp. 1537–1541.
14. Girshick R. Fast R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision; 2015. pp. 1440–1448.
15. Zheng C, Chen P, Pang J, et al. A mango picking vision algorithm on instance segmentation and key point detection from RGB images in an open orchard. *Biosystems Engineering* 2021; 206: 32–54. doi: 10.1016/j.biosystemseng.2021.03.012
16. Qiang B, Chen R, Zhou M, et al. Convolutional neural networks-based object detection algorithm by jointing semantic segmentation for images. *Sensors* 2020; 20(18): 5080. doi: 10.3390/s20185080
17. Guo Y, Liu Y, Georgiou T, et al. A review of semantic segmentation using deep neural networks. *International Journal of Multimedia Information Retrieval* 2017; 7(2): 87–93. doi: 10.1007/s13735-017-0141-z
18. Kasarla T, Nagendar G, Hegde GM, et al. Region-based active learning for efficient labeling in semantic segmentation. In: Proceedings of the 2019 IEEE Winter Conference on Applications of Computer Vision (WACV); 2019. pp. 1109–1117.
19. Caesar H, Uijlings J, Ferrari V. Region-based semantic segmentation with end-to-end training. In: Proceedings of the Computer Vision—ECCV 2016: 14th European Conference; 11–14 October 2016; Amsterdam, The Netherlands. pp. 381–397.
20. Sun W, Wang R. Fully convolutional networks for semantic segmentation of very high resolution remotely sensed images combined with DSM. *IEEE Geoscience and Remote Sensing Letters* 2018; 15(3): 474–478. doi: 10.1109/lgrs.2018.2795531
21. Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 7–12 June 2015. pp. 3431–3440.
22. Hao S, Zhou Y, Guo Y. A brief survey on semantic segmentation with deep learning. *Neurocomputing* 2020; 406: 302–321. doi: 10.1016/j.neucom.2019.11.118
23. Pathak D, Krahenbuhl P, Darrell T. Constrained convolutional neural networks for weakly supervised segmentation. In: Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV); 07–13 December 2015. pp. 1796–1804.
24. Papandreou G, Chen LC, Murphy KP, Yuille AL. Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In: Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV); 7–13 December 2015. pp. 1742–1750.
25. Alzahrani N, Al-Baity HH. Object recognition system for the visually impaired: A deep learning approach using Arabic annotation. *Electronics* 2023; 12(3): 541. doi: 10.3390/electronics12030541
26. Tamulionis M, Sledevič T, Abromavičius V, et al. Finding the least motion-blurred image by reusing early features of object detection network. *Applied Sciences* 2023; 13(3): 1264. doi: 10.3390/app13031264
27. Wu J, Shen T, Wang Q, et al. Local adaptive illumination-driven input-level fusion for infrared and visible object detection. *Remote Sensing* 2023; 15(3): 660. doi: 10.3390/rs15030660
28. Zhu Y, Xu R, An H, et al. Anti-noise 3D object detection of multimodal feature attention fusion based on PV-RCNN. *Sensors* 2022; 23(1): 233. doi: 10.3390/s23010233
29. Fang X, Jiang M, Zhu J, et al. M2RNet: Multi-modal and multi-scale refined network for RGB-D salient object detection. *Pattern Recognition* 2023; 135: 109139. doi: 10.1016/j.patcog.2022.109139
30. Dharmik RC, Chavhan S, Sathe SR. Deep learning based missing object detection and person identification: An application for smart CCTV. *3C Tecnología_Glosas de Innovación Aplicadas a la Pyme* 2022; 11(2): 51–57. doi: 10.17993/3ctecno.2022.v11n2e42.51-57
31. Nguyen T, Hua BS, Le N. 3D-UCaps: 3D capsules unet for volumetric image segmentation. In: Proceedings of the Medical Image Computing and Computer Assisted Intervention—MICCAI 2021: 24th International Conference; 27 September–1 October 2021; Strasbourg, France. pp. 548–558.
32. Liu Z, Shi P, Qi H, et al. D-S Augmentation: Density-semantics augmentation for 3-D object detection. *IEEE Sensors Journal* 2023; 23(3): 2760–2772. doi: 10.1109/jsen.2022.3231882
33. Mahayuddin ZR, Saif AFMS. Moving object detection using semantic convolutional features. *Journal of Information System and Technology Management* 2022; 7(29): 24–41. doi: 10.35631/jistm.729003
34. Xia T, Yang J, Chen L. Automated semantic segmentation of bridge point cloud based on local descriptor and machine learning. *Automation in Construction* 2022; 133: 103992. doi: 10.1016/j.autcon.2021.103992
35. Sun X, Chen C, Wang X, et al. Gaussian dynamic convolution for efficient single-image segmentation. *IEEE Transactions on Circuits and Systems for Video Technology* 2022; 32(5): 2937–2948. doi: 10.1109/tcsvt.2021.3096814
36. Rachmatullah MN, Nurmaini S, Sapitri AI, et al. Convolutional neural network for semantic segmentation of fetal echocardiography based on four-chamber view. *Bulletin of Electrical Engineering and Informatics* 2021; 10(4): 1987–1996. doi: 10.11591/eei.v10i4.3060
37. Shan J, Li X, Jia S, et al. Semantic segmentation based on deep convolution neural network. *Journal of Physics: Conference Series* 2018; 1069: 012169. doi: 10.1088/1742-6596/1069/1/012169

38. Zhang Y, Chu J, Leng L, et al. Mask-Refined R-CNN: A network for refining object details in instance segmentation. *Sensors* 2020; 20(4): 1010. doi: 10.3390/s20041010
39. Liu W, Anguelov D, Erhan D, et al. SSD: Single shot multibox detector. In: Proceedings of the Computer Vision—ECCV 2016: 14th European Conference; 11–14 October 2016; Amsterdam, the Netherlands. pp. 21–37.
40. Ren S, He K, Girshick R, Sun J. Faster R-CNN: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems* 2015; 28.
41. Liu Y, Li J, Wang Y, et al. Refined segmentation R-CNN: A two-stage convolutional neural network for punctate white matter lesion segmentation in preterm infants. In: Proceedings of the Medical Image Computing and Computer Assisted Intervention—MICCAI 2019: 22nd International Conference; 13–17 October 2019; Shenzhen, China. pp. 193–201.
42. Bazgir O, Zhang R, Dhruba SR, et al. Representation of features as images with neighborhood dependencies for compatibility with convolutional neural networks. *Nature Communications* 2020; 11(1). doi: 10.1038/s41467-020-18197-y
43. Parmar Y, Natarajan S, Sobha G. DeepRange: Deep-learning-based object detection and ranging in autonomous driving. *IET Intelligent Transport Systems* 2019; 13(8): 1256–1264. doi: 10.1049/iet-its.2018.5144