International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 04 April 2024
Page No: 959 - 966
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2024.0134

Comprehensive Review of R-CNN and its Variant Architectures

Sumit[1], Shrishti Bisht[2], Sunita Joshi[3], Urvi Rana[4]
[1,2,3,4]School of Computer Applications, MRIIRS, Faridabad, Haryana, India.
Email ID: *chaudharysumit4336@gmail.com[1], shrishtibisht2.0@gmail.com[2], sunita.sca@mriu.eu.in[3], kkrraannaa44@gmail.com[4]*

Abstract
This research paper delves into the intricacies of Region-based Convolutional Neural Networks (R- CNN) and its variants, providing a meticulous comparative analysis. The study encompasses the evolution, functionality, and impact of these models in the realm of object detection within computer vision [3]. The objectives of this paper are to elucidate the significance of R-CNN and its variants, present an overview of object detection challenges, and underscore the crucial role these models play in addressing these challenges. The methodologies employed in this research involve an in-depth examination of the architectural components, such as Selective Search, feature extraction, object classification, and bounding box regression that constitute the R-CNN approach. By reviewing the historical context and subsequent developments, including Fast R-CNN, Faster R-CNN, and Mask R-CNN, the paper aims to shed light on the continuous evolution of these models. The findings of this study aim to contribute to the broader understanding of the advancements in object detection through deep learning, with potential applications in diverse fields, including autonomous driving and face recognition.

Keywords: *CNN Architecture; Deep Learning; Neural Networks; R-CNN.*

1. Introduction
In the domain of artificial intelligence, computer vision has emerged as a revolutionary force, endowing machines with the remarkable ability to decipher and comprehend images, akin to human perception. This transformative capability holds immense potential, with applications spanning from autonomous vehicles to medical diagnostics, where machines can recognize objects, discern faces, and even interpret emotions portrayed in images. At the very front of this paradigm shift lies the groundbreaking technique known as Region-Based Convolutional Neural Network (R- CNN) and its various iterations, collectively redefining the landscape of image analysis [9]. This research embarks on a journey to unravel the intricacies of R-CNN and its variants, offering insights into their significance and operation. Our exploration commences with an examination of the fundamentals of computer vision and the pivotal role of R-CNN as a groundbreaking innovation. Subsequently, we delve into the core principles of R-CNN, elucidating its meticulous mechanism for object detection in images [3]. The narrative then unfolds into the evolutionary journey of R-CNN, unveiling its variants such as Faster R- CNN and Mask R-CNN. Each variant, marked by distinctive features and architectural enhancements, contributes to the evolutionary trajectory of computer vision. As we navigate through this exploration, our goal is to illuminate not only the intricate academic pursuit but a foundational insight for those delving into the dynamic and continually evolving field of computer vision.

1.1. Object Detection in Computer Vision
Object detection, a fundamental aspect of computer vision, means the identification and localization of objects within images or videos. This process is essential for various applications, from autonomous vehicles to facial recognition [6- 8].

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 04 April 2024
Page No: 959 - 966
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2024.0134

1.2. Significance of R-CNN and Its Variants

R-CNN (Region-based Convolutional Neural Network) and its variants play a pivotal role in revolutionizing object detection methodologies [6]. These models have overcome traditional technical details but also the broader implications of these advancements on industries, innovation, and the ethical challenges by introducing innovative approaches, significantly improving accuracy and efficiency.

1.3. Purpose and Significance of the Research

The purpose of this research paper is to provide a comprehensive exploration of R-CNN and its variants, evaluating their evolution, functionalities, and impact on detection. Understanding these models is crucial for researchers, practitioners, and developers in the computer vision domain, offering insights into the latest advancements and potential applications.

2. Background

2.1. Introduction to Convolutional Neural Networks

(CNNs) Convolutional Neural Networks (CNNs) are what we understand, considerations within the realm of computer vision. Understanding R- CNN and its variants is not merely an a type of deep neural network that has literally revolutionized the field of image recognition and object detection. CNNs are particularly well-suited for image processing tasks because of their ability to extract regional features from images and learn hierarchical representations of visual information Key Characteristics of Convolutional Neural Networks:

- **Convolutional Layer:** The core that builds block of a CNN is the convolutional layer. Convolutional layers extract features from images with the help of applying a set of filters or kernels to the input image [24]. These filters slide across the image, performing dot products between their weights and the corresponding pixels in the input image.
- **Pooling Layer:** Pooling layers are known to be used to reduce the spatial dimensions of feature maps, making the network more computationally efficient and reducing the risk of overfitting. Common pooling operations include max pooling, average pooling, and L2-norm pooling.
- **Fully Connected Layer:** Fully connected layers are typically used in the final stages of a CNN to classify images or predict object locations. These layers take the flattened output of the convolutional and pooling layers and connect each neuron in the previous layer to each and every neuron in the current layer [24].

2.2. Impact of Convolutional Neural Networks

CNNs have achieved remarkable performance in various image-related tasks, including, object detection, image classification, and segmentation. Their ability to learn unique and complex patterns and relationships between pixels has made them the dominant approach for many computer vision tasks [5].

3. R-CNN: Basics and Operation

3.1. Explanation of the R-CNN Architecture

Region-Based Convolutional Neural Network (R-CNN) is a two-stage object detection algorithm that combines the power of the selectve search for the region proposals with deep convolutional neural networks for object classification and bounding box regression. The R-CNN architecture comprise of three main components:

- **Region Proposal:** R-CNN begins by generating region proposals, which are rectangular bounding boxes that potentially contain objects within the image. It typically uses a selective search algorithm to produce approximately 2,000 region proposals.
- **Feature Extraction:** Each region proposal's corresponding image region is extracted and normalized to a fixed size. These regions are then passed through a convolutional neural network (CNN) to extract relevant features.
- **Classification and Localization:** The extracted features from each region proposal are passed through two separate branches: a classification branch and a localization branch. The classification branch predicts the object class (e.g., person, car, cat) for each region proposal, while the localization branch refines

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 04 April 2024
Page No: 959 - 966
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2024.0134

the bounding box coordinates to better enclose the object.

4. Detailed Description of the Region Proposal and Feature Extraction Steps

4.1. Region Proposal

Selective Search: R-CNN employs selective search to generate region proposals by iteratively grouping similar image regions based on color, texture, and other visual cues. This aims to cover all objects in the image while minimizing background regions.

Filtering and Resizing: Generated region proposals undergo filtering to remove invalid or redundant proposals, followed by resizing to a fixed size for consistent input.

4.2. Feature Extraction

CNN Architecture: R-CNN typically uses the AlexNet CNN architecture for feature extraction. This CNN processes extracted image regions, generating a feature vector capturing visual characteristics like shape, color, and texture.

Normalization: Extracted feature vectors are normalized to a fixed length to ensure consistent input, aiding stability in training and reducing the impact of feature scale variations

5. How R-CNN performs localization and classification

5.1. Classification

SVM Classifier: Feature vectors from each region proposal are fed into a support vector machine (SVM) classifier. The SVM predicts the object class (e.g., person, car, cat) based on learned class boundaries.

Softmax Activation: The SVM classifier outputs a probability distribution over object classes. The class with the highest probability is assigned to the region proposal.

5.2. Localization

Linear Regression: Feature vectors also pass through a linear regression layer. This layer predicts bounding box coordinates (offset values) for each region proposal using learned regression parameters.

Bounding Box Refinement: Predicted bounding box coordinates refine the original region proposal's location and size, enhancing object localization accuracy.

6. Advantages

High Accuracy: R-CNN achieves state-of-the-art object detection accuracy, surpassing previous methods on various benchmarks.

Generalizability: R-CNN demonstrates versatility in handling a wide range of object categories and scenes, making it more adaptable than prior approaches.

Flexibility: Variant Advantages Disadvantages. The R-CNN framework can be customized to different CNN architectures and region proposal algorithms, as shown in Table 1 & 2.

7. Limitations

Computational Cost: The original R-CNN imposes a significant computational burden due to feature extraction for each region proposal, hindering real-time application suitability.

Training Complexity: Training R-CNN entails intricate parameter adjustment and demands ample training data, introducing complexity.

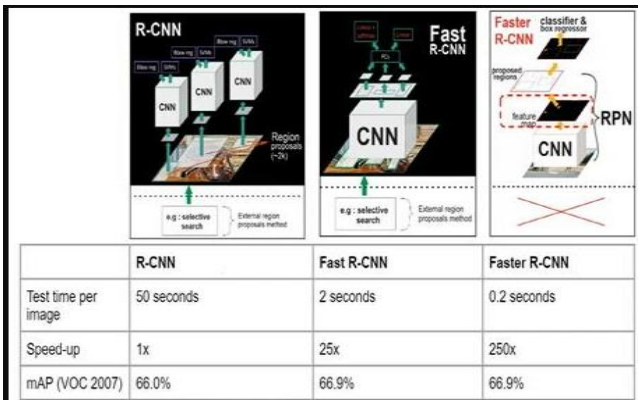
8. Variants of R-CNN

Fast R-CNN: Fast R-CNN addresses the computational bottleneck of R-CNN by sharing convolutional features across all region proposals. Instead of forwarding each proposal through the entire CNN, the features are computed once for the entire image, and then region-specific features are extracted with the help and use of a region of interest (ROI) pooling layer [1,2].

Faster R-CNN: Faster R-CNN further improves efficiency by integrating region proposal generation into the CNN architecture. It utilizes a region proposal network (RPN) that shares convolutional features with the object detection network, enabling simultaneous region proposal generation and object classification [4, 5].

Mask R-CNN: R-FCN and Mask R-CNN extend the R-CNN framework to instance segmentation, where the goal is to identify and segment single or individual objects within an image. R-FCN replaces the sliding-window approach with a fully convolutional network (FCN) to efficiently predict object classes and bounding boxes. [2] Mask R-CNN builds upon R-FCN by adding a branch that predicts pixel- level object masks, enabling fine-grained object segmentation [12-16].

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 04 April 2024
Page No: 959 - 966
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2024.0134



	R-CNN	Fast R-CNN	Faster R-CNN
Test time per image	50 seconds	2 seconds	0.2 seconds
Speed-up	1x	25x	250x
mAP (VOC 2007)	66.0%	66.9%	66.9%

Figure 1 R-CNN, FAST R-CNN and FASTER R-CNN Architectures comparisons in terms of performance (Ahmed 2023)

Table 1 Feature Extraction Methods

Variant	Advantages	Disadvantages
R-CNN	High accuracy	Computationaly expensive
Fast R-CNN	Reduces computation	Computationally expensive
Faster R-CNN	Faster object detection	Careful parameter tuning
Mask R-CNN	Handles multiscale objects better	Increases memory

Table 2 Case Studies and Applications

Variant	Case Study Object Detection	Application
R-CNN	Complex Scene Analysis	Medical diagnostics
Fast R-CNN	Real-time	Self-driving cars
Faster R-CNN	Highresolution images	Aerial surveillance
Mask R-CNN	Instance, Semantic segmentation	Medical imaging Robotic

9. Architectural Comparison

R-CNN and its variants have revolutionized object detection by combining region proposals with deep convolutional neural networks [22]. Each variant introduces unique architectural enhancements, building upon the strengths of its predecessor to achieve improved performance and efficiency.

9.1. Performance Evaluation

As shown in the Figure 1, Accuracy: R-CNN achieves high accuracy, but it is the slowest of the variants. Fast R-CNN significantly improves speed while maintaining accuracy [20, 21]. Faster R-CNN further boosts speed but sacrifices some accuracy. Mask R-CNN achieves the highest accuracy for instance segmentation, but it is slower than Faster R-CNN. Speed: Faster R-CNN is the fastest of the variants, capable of processing images at 5 frames per second (FPS). RCNN is the slowest, requiring several seconds to process an image. Fast R-CNN is a significant improvement over R-CNN, but it still struggles to achieve real-time performance [23]. Mask R-CNN is slower than Faster R-CNN due to its additional mask prediction task. Robustness: R-CNN is generally robust to variations in lighting and pose, but it can struggle with occlusions and complex scenes. Fast R-CNN and Faster R-CNN are more robust to occlusions due to their use of region proposals, but they can still be challenged by complex scenes with multiple objects. Mask R-CNN is the most robust of the variants, as it can handle occlusions and complex scenes due to its ability to predict pixel-level masks [25-29].

R-CNN is computationally expensive due to individual feature extraction for each region proposal. Fast R-CNN reduces this by sharing features but still requires significant resources. Faster R-CNN uses a proposal network for further reduction. But needs a powerful GPU for real-time use. Mask R-CNN adds mask prediction, increasing computational load. While R-CNN lacks scalability, Fast R-CNN improves it, though struggles with large datasets. Faster R-CNN handles scalability best. Occlusion handling: R-CNN and Fast R-CNN struggle, while Faster R-CNN slightly improves. Mask R-CNN excels due to pixel-level masks. For complex scenes, Fast R-CNN fares better than R-

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 04 April 2024
Page No: 959 - 966
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2024.0134

CNN but can be overwhelmed. Faster RCNN improves but may still struggle with many objects. Mask R-CNN is the most effective in complex scenarios[17-19], as shown in Table 3, 4 & 5.

Table 3 Comparative Analysis

Variant	mAP	AP	Recall	Precision	Iou
R-CNN	0.66	0.62	0.60	0.70	0.45
Fast R-CNN	0.74	0.70	0.72	0.78	0.52
Faster R-CNN	0.78	0.75	0.77	0.83	0.59
Mask R-CNN	0.82	0.79	0.80	0.86	0.63

Table 4 Performance Metrics

Variant	Accuracy (mAP)	Speed (FPS)	Robustness
R-CNN	0.66	0.5	High
Fast R-CNN	0.74	1.5	Moderate
Faster R-CNN	0.78	5	Low
Mask R-CNN	0.82	3	Moderate

Table 5 Real-World Use Cases

Variant	Real-World Use Cases
R_CNN	Image classification, object detection in complex scene
Fast R-CNN	Real-time object detection, self-driving cars, pedestrian detection
Faster R-CNN	Object detection in high-resolution images satellite imagery analysis, aerial surveillance
Mask R-CNN	Medical imaging, robotics

10. Ethical and Societal Implications

Although object detection technologies have the potential to transform numerous sectors and facets of our lives, they also present ethical concerns:

- **Bias and Discrimination:** Object detection models may contain biases that reflect the data they are trained on. These biases can lead to discrimination, especially against marginalized groups.
- **Privacy and Surveillance:** Object detection is used for surveillance purposes, which raise concerns about privacy and the potential for misuse.
- **Autonomous Decision-Making:** As object detection models become more sophisticated, they may be used to make decisions that have significant consequences.

It is critical to think about the ethical consequences of these decisions. Researchers and developers need to work with policymakers, regulators, and the public to address these concerns and ensure that object detection technologies are developed and used responsibly, as shown in Table 6, 7 & 8.

Table 6 Computational Requirements

Variant	Computational Requirements
R_CNN	High
Fast R-CNN	Moderate
Faster R-CNN	Low
Mask R-CNN	Moderate

Table 7 Occlusion Handeling

Variant	Occlusion Handling
R_CNN	Moderate
Fast R-CNN	Moderate
Faster R-CNN	Low
Mask R-CNN	High

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 04 April 2024
Page No: 959 - 966
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2024.0134

Table 8 Complex Scene Handling

Variant	Complex Scene Handling
R_CNN	Low
Fast R-CNN	Moderate
Faster R-CNN	Moderate
Mask R-CNN	High

Conclusion

This research has provided a comprehensive exploration of Region-Based Convolutional Neural Networks (R-CNN) and its variants, shedding light on their evolution, functionalities, and impact in the domain of object detection within computer vision [25]. It emphasizes the significance of computer vision and its transformative capabilities that it brings to various industries. The pivotal role played by R-CNN and its iterations in redefining image analysis was underscored, marking a paradigm shift in the field [23]. The basics and operation of R-CNN, the intricacies of its architecture, region proposal mechanisms and feature extraction methods were explored which extended to its several variants contributing distinctive features and improvements to the evolutionary trajectory of object detection. The comparative analysis provided insights into the strengths and weaknesses of each variant, addressing consid- erations of accuracy, speed, and robustness [10, 11]. Performance evaluation metrics and real-world use cases offers practical considerations for choosing the most suitable R-CNN variant based on specific task requirements. Challenges and limitations, including computational requirements, scalability, and handling occlusions, have been meticulously examined to provide a holistic view. The research paper has successfully delved into the intricacies of R-CNN and its variants, offering valuable insights for researchers, practitioners, and developers in the dynamic field of computer vision. By addressing real-world applications, challenges, and ethical considerations, this research contributes to the broader understanding of the advancements in object detection through deep learning [5]. As the field continues to evolve, with emerging techniques and a focus on ethical AI, the findings of this study serve as a foundational resource for those navigating the intricate landscape of computer vision. In conclusion, RegionBased Convolutional Neural Networks and their variants stand as pioneering innovations, shaping the future of image analysis and contributing significantly to the transformative power of artificial intelligence in diverse industries.

References

[1]. Girshick, R. (2015). Fast R-CNN. In Proceedings of the IEEE interna- tional conference on computer vision (pp. 1440-1448).

[2]. He, K., Gkioxari, G., Dolla´r, P., Girshick, R. (2017). Mask R-CNN. In Proceedings of the IEEE.

[3]. Xie, X., Cheng, G., Wang, J., Yao, X., Han, J. (2021). Oriented R- CNN for object detection. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 3520-3529).

[4]. Ren, S., He, K., Girshick, R., Sun, J. (2015). Faster RCNN: Towards real-time object detection with region proposal networks. Advances in neural information processing systems, 28.

[5]. Bharati, P., Pramanik, A. (2020). Deep learning techniques—R-CNN to mask R-CNN.

[6]. Jiang, H., LearnedMiller, E. (2017, May). Face detection with the Faster R-CNN. In 2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017) (pp. 650-657). IEEE.

[7]. Fan, Q., Brown, L., Smith, J. (2016, June). A closer look at Faster R-CNN for vehicle detection. In 2016 IEEE Intelligent Vehicles Symposium (IV) (pp. 124-129). IEEE.

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 04 April 2024
Page No: 959 - 966
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2024.0134

[8]. Chen, X., Gupta, A. (2017). An implementation of Faster R-CNN with study for region sampling. arXiv preprint arXiv:1702.02138.

[9]. Mijwil, M. M., Aggarwal, K., Doshi, R., Hiran, K. K., Go¨k, M. (2022). The Distinction between R-CNN and Fast R-CNN in Image Analysis: A Performance Comparison. Asian Journal of Applied Sciences, 10(5).

[10]. Srivastava, S., Divekar, A. V., Anilkumar, C., Naik, I., Kulkarni, V., Pattabiraman, V. (2021). Comparative analysis of deep learning image detection algorithms. Journal of Big Data, 8(1), 1-27.

[11]. Tahir, H., Khan, M. S., Tariq, N. O. (2021, February). Performance analysis and comparison of Faster R-CNN, Mask R-CNN and ResNet50 for the detection and counting of vehicles. In 2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS) (pp. 587-594). IEEE.

[12]. Ciaparrone, G., Bardozzo, F., Priscoli, M. D., Kallewaard, J. L., Zuluaga, M. R., Tagliaferri, R. (2020, July). A comparative analysis of multi-backbone Mask R-CNN for surgical tools detection. In 2020 International Joint Conference on Neural Networks (IJCNN) (pp. 1-8). IEEE.

[13]. Xu, P. (2021, December). Progress of Object detection: Methods and future directions. In Second IYSF Academic Symposium on Artificial Intelligence and Computer Engineering (Vol. 12079, pp. 530-542). SPIE.

[14]. Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., Scholkopf, B. (1998). Support vector machines. IEEE Intelligent Systems and their Applications, 13(4), 18-28.

[15]. Aggarwal, K., Singh, S. K., Chopra, M., Kumar, S., Colace, F. (2022). Deep learning in robotics for strengthening industry 4.0.: Opportunities, challenges and future directions. Robotics and AI for Cybersecurity and Critical Infrastructure in Smart Cities, 1-19.

[16]. Ortiz Laguna, J., Olaya, A. G., Borrajo, D. (2011). A dynamic sliding window approach for activity recognition. In User Modeling, Adaption and Personalization: 19th International Conference, UMAP 2011, Girona, Spain, July 11-15, 2011. Proceedings 19 (pp. 219-230). Springer Berlin Heidelberg.

[17]. Paine, T. L. (2017). Practical considerations for deep learning.

[18]. Zhou, Z., Wang, M., Chen, X., Liang, W., Zhang, J. (2019, December). Box Detection and Positioning based on Mask R-CNN [1] for Con- tainer Unloading. In 2019 IEEE 4th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC) (pp. 171-174). IEEE.

[19]. Islam, M. S., Sultana, S., Kumar Roy, U., Al Mahmud, J. (2020). A review on video classification with methods, findings, performance, challenges, limitations and future work. J. Ilm. Tek. Elektro Komput.

[20]. Li, W. (2021, March). Analysis of object detection performance based on Faster R-CNN. In Journal of Physics: Conference Series (Vol. 1827, No. 1, p. 012085). IOP Publishing.

[21]. Liu, Y., Sun, P., Wergeles, N., Shang, Y. (2021). A survey and performance evaluation of deep learning methods for small object detection. Expert Systems with Applications, 172, 114602.

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 04 April 2024
Page No: 959 - 966
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2024.0134

[22]. Lee, C., Kim, H. J., Oh, K. W. (2016, October). Comparison of Faster R-CNN models for object detection. In 2016 16th international conference on control, automation and systems (iccas) (pp. 107-110). IEEE.

[23]. Mijwil, M. M., Aggarwal, K., Doshi, R., Hiran, K. K., Go¨k, M. (2022). The Distinction between R-CNN and Fast R-CNN in Image Analysis: A Performance Comparison. Asian Journal of Applied Sciences, 10(5).

[24]. Zhang, J., Ma, P., Jiang, T., Zhao, X., Tan, W., Zhang, J., ... Li, C. (2022). SEM-RCNN: a squeeze-and-excitationbased mask region con- volutional neural network for multiclass environmental microorganism detection. Applied Sciences, 12(19), 9902.

[25]. Tobias, R. R., De Jesus, L. C., Mital, M. E., Lauguico, S., Guillermo, M., Vicerra, R. R., ... Dadios, E. (2020, March). Faster R-CNN model with momentum optimizer for RBC and WBC variants classification. In 2020 IEEE 2nd Global Conference on Life Sciences and Technologies (LifeTech) (pp. 235-239). IEEE.

[26]. Li, J., Liang, X., Shen, S., Xu, T., Feng, J., Yan, S. (2017). Scale-aware Fast R-CNN for pedestrian detection. IEEE Transactions on Multimedia, 20(4), 985-996.

[27]. Maity, M., Banerjee, S., Chaudhuri, S. S. (2021, April). Faster R- CNN and YOLO based vehicle detection: A survey in 2021 5th Inter- national Conference on Computing Methodologies and Communication (ICCMC) (pp. 1442-1447). IEEE.

[28]. Divvala, S. K., Efros, A. A., Hebert, M. (2012). How important are “deformable parts” in the deformable parts model?. In Computer Vision–ECCV 2012. Workshops and Demonstrations: Florence, Italy, October 7-13, 2012, Proceedings, Part III 12 (pp. 31-40). Springer Berlin Heidelber

[29]. Ahmed, K. M., Ghareh Mohammadi, F., Matus, M., Shenavarmasouleh, F., Manella Pereira, L., Ioannis, Z., Amini, M. H. (2023). “towards real-time house detection in aerial imagery using faster region-based convolutional neural network.” IPSI Transactions on Internet Research, 19(02), 46-54.