Two Stage Traffic Sign Detection and Recognition based on K-Means and Mask RCNN Algorithm

Parv Dave[1], Meet Diwan[2], Meet Raval[3]

[1-2]G H Patel College of Engineering & Technology
[3]Dhirubhai Ambani Institute of Information and Communication Technology

---

Abstract— Sign board identification is critical in machine learning applications such as video surveillance and content-based visual information retrieval. Previous studies on this problem have mostly focused on application-specific sign boards. To detect these sign boards, several special qualities such as hue, size, form, and symmetrical must be employed. In this research, we propose a technique for detecting sign boards in images or videos using Mask RCNN. Experiments on over a hundred photos clearly illustrate the new algorithms' usefulness. Image segmentation is a critical issue in image processing and computer vision, with several applications including scene interpretation, medical picture analysis, robotic perception, video surveillance, augmented reality, and image compression. Several picture segmentation techniques have been developed in the literature. Due to the success of deep learning models in a variety of vision applications, there has recently been a significant amount of effort directed towards creating picture segmentation algorithms utilising deep learning models.

Keywords—classification of images, data mining, decision tree, object recognition, satellite images, convolutional neural networks (CNNs); end-to-end detection; transfer learning; remote sensing

I. INTRODUCTION

Image Segmentation is a critical component in many visual comprehension systems. It entails dividing a picture into several segments or objects. To name a few, segmentation is essential in a wide range of applications such as medical image analysis, driverless cars, video surveillance, and augmented reality. In the literature, numerous image segmentation algorithms have been developed, ranging from the earliest methods, such as thresholding, histogram-based bundling, region growing, k-means clustering, and watersheds, to more advanced algorithms, such as active contours, graph cuts, conditional and Markov random fields, and sparsity-based methods.

Signs give a series of cautions regarding the route. They keep traffic moving by assisting passengers in achieving their destinations and informing them of arrival, departure, and turn locations ahead of time. Signs are strategically positioned to guarantee the safety. They also provide information on when and where drivers may or may not turn. We designed a method for sign detection and recognition in this research, as well as a method for extracting a road sign from a natural complex image and processing it. It is used in such a way that it assists drivers in making quick judgments. Factors such as changeable weather, changing light directions, and varied light intensity make sign recognition difficult in real-time settings. Noise, partial or absolute underexposure, partial or entire overexposure, and considerable fluctuations in colour saturation, broad range of viewing angles, view depth, and shape/colour deformations of traffic signs all have an impact on the machine's dependability. The planned architecture is divided into three sections. The first is picture pre-processing, which involves quantifying the dataset's input files, determining the input size for learning, and resizing the information for the learning phase. During the recognition phase, the suggested method categorizes the symbol on the sign board. In the second phase, a Convolutional Neural Network method is utilised to achieve this, and the third phase deals with text-to-speech translation, with the recognised sign from the second phase delivered in audio format. Mask R-CNN expands Faster R-CNN by adding a branch for predicting segmentation masks for each Region of Interest, in addition to the current branches for classification and bounding box regression. The mask branch is a tiny FCN that is applied to each RoI and predicts a segmentation mask pixel-by-pixel. Mask R-CNN is easy to develop and train because to the Faster R-CNN framework, which allows for a variety of configurable architectural designs. Furthermore, the mask branch adds just a minor computational burden, allowing for a quick system and rapid experimentation.

Despite appearing to be a modest improvement, RoIAlign has a significant impact: it increases mask accuracy by 10% to 50%, with larger benefits under tougher localization measures. Second, we discovered that decoupling mask and class prediction is critical: we predict a binary mask for each class individually, without competition between classes, and rely on the network's RoI classification branch to forecast the category. In contrast, FCNs typically do per-pixel multi-class categorization, which combines segmentation and classification and, according to our findings, performs badly for segmentation. The goal of object extraction is to extract the Region of Interest of a given image based on the identified position of an item of interest. Object extraction has been widely employed in varied applications such as real-time monitoring, robot navigation, and target search as a common issue in the field of image processing and computer vision. Despite significant recent advances in object extraction, there are still many unresolved issues. For example, owing to the interference generated by diverse views and settings from where the picture was acquired, it is currently impossible to ensure the reliability and resilience of the object model obtained from an image.

Despite the effectiveness of the YOLO v3 framework, the algorithm's related complicated pipelines limit its wider applicability due to high computation time and prohibitive hardware requirements. Mask RCNN implements object detection as a single regression issue, resulting in effective performance when compared to state-of-the-art approaches. The upgraded classifier network and independent logic classifier were utilised to enhance the Mask RCNN algorithm's properties. Faster RCNN has been more accurate, with the extra bonus of running eight times faster. Because of these characteristics, Mask RCNN has become the most used method for picture semantic object recognition. Instance segmentation is difficult because it involves accurate recognition of all objects in a picture as well as exact segmentation of each instance. It thus combines elements from the classical computer vision tasks of object detection, in which the goal is to classify individual objects and localise each using a bounding box, and semantic segmentation, in which the goal is to classify each pixel into a fixed set of categories without distinguishing object instances. Given this, one could assume a sophisticated procedure to be necessary to attain satisfactory results. However, we demonstrate that a system that is surprisingly simple, flexible, and quick can outperform previous state-of-the-art instance segmentation outcomes Fortunately, the emergence of Convolutional Neural Networks has allowed for the creation of more sophisticated and precise approaches for object detection and extraction. Object semantic detection based on CNN, the fundamental technology of object extraction, has been extensively explored due to its potential for improving robot intelligence and autonomy, a field in which enhanced sensing and object perception mechanisms are required. Despite improvements in object detection accuracy, detection speed and identification of objects with complicated geometries still need to be improved. Faster R-CNN has been developed in this manner, giving rise to Mask R-CNN by adding the capacity to forecast segmentation masks on each Region of Interest constituting the picture. It is feasible to accomplish concurrent object detection and extraction with such contributions.

Over a short period of time, the vision community has quickly improved object detection and semantic segmentation outcomes. These advancements have been mostly driven by powerful baseline systems, such as the Fast/Faster RCNN and Fully Convolutional Network frameworks for object identification and semantic segmentation. These approaches are theoretically simple and provide flexibility and resilience, as well as quick training and inference times. Our objective in this effort is to provide a comparable enabling structure, such as segmentation.

II. LITERATURE SURVEY

Studies on intelligent cars that drive themselves in urban environments are getting increasingly prominent these days. Detecting traffic lights in real-world driving situations is a difficult problem. Despite the efforts of numerous academics to offer the reliability necessary by autonomous cars to safely pass-through junctions, most of these solutions are, to varying degrees, flawed in real-world driving environments. Traffic symbols have various unique characteristics that may be used to detect and identify them.

They are created in precise colours and forms, with the word or symbol standing out against the backdrop. To identify traffic signals in early works, a vision-based approach is utilised. However, this technique necessitates the installation of a camera in a fixed location near traffic lights, and these technologies are incapable of meeting the real-time processing requirement. As a result, these traffic light detection technologies do not apply to intelligent cars. In the case of traffic sign detection, most systems employ colour information to separate pictures. In scenarios with bright illumination, low lighting, or terrible weather conditions, the performance of colour-based traffic sign identification is frequently degraded. The great majority of existing systems rely on manually labelling real photos, which is a time-consuming and error-prone procedure. Traffic symbol information, such as shape and colour, may be utilised to categorise traffic symbols; nevertheless, there are various issues that might impede successful detection and identification of traffic signals. Variations in light occlusion of signs, motion blur, and weather-worn degeneration of signs are among these issues. The road landscape is also rather congested, with several bold geometric forms that may easily be misidentified as traffic signs. Accuracy is critical since even one incorrectly categorised or recognised indication might have a negative influence on the driver. The most important step in any type of research is to do a literature review. This approach would allow us to detect any holes or faults in the present framework, allowing us to try to find a way around the restrictions of the current system. In this section, we briefly examine comparable work on traffic sign detection, identification, and recognition.

Deep Learning for Large Scale Traffic-Sign Detection and Recognition by DomenTabernik; DanijelSkočaj used CNN, specifically the R-CNN mask, is utilised for traffic sign detection and recognition. They created a new data set called DFG traffic-sign to have low inter-class variability and high intra-class variability. The disadvantages of these were that instances involving missed traffic signs were not considered.

The Speed Limit Road Signs Recognition Using Hough Transformation and Multi-Class Svm by Ivona Matoš; Zdravko Krpić; Krešimir Romić used SVM is utilised for classification, while the HOG descriptor is employed for feature extraction. The system's limits were that images with a lot of noise were processed successfully, and up to 95 percent performance was attained.

Traffic Sign Detection and Recognition using a CNN Ensemble by AashrithVennelakanti, Smriti Shreya, ResmiRajendran, Debasis Sarkar, Deepak Muddegowda, PhanishHanagal used for color-based detection, the Hue Saturation Value (HSV) colour space is employed instead of RGB, while the Douglas Peucker algorithm is used for shape-based detection. The restriction was that while good accuracy was obtained, only triangular and circular forms were considered for detection.

III. MEHODOLOGY

K-MEANS SEGMENTATION ALGORITHM:

The breakdown of an image into many non-overlapping meaningful sections with or without the same features is referred to as image segmentation. In digital image processing, image segmentation is a key algorithm. As a result, the accuracy of segmentation has a direct impact on the effectiveness of any application that requires such picture information.

The segmentation process in traditional image segmentation algorithms is based on the three image properties listed below:

1) The threshold,

2) The border, and

3) The territory

The algorithms of classic picture segmentation approaches are based on cluster analysis theory, which people utilise while learning to identify items by constantly altering the subconscious clustering pattern. The K-means segmentation approach is used in the suggested picture segmentation because to the great efficiency of the conventional K-means cluster analysis in large-scale data. The main principle behind the K-means method is to split the provided data into groups based on the clustering number K, with the points in each cluster being the closest to each other. The K-means algorithm is a common formulation of the Minkowski Distance function-based clustering approach.

The typical K-means method is described in four steps:

Step 1: Using the provided picture data collection, randomly establish K example cluster centres.

Step 2: To establish a cluster, assign all picture data points that are closest to a specific example cluster centre using some distance function.

Step 3: Using all data points allocated to the associated cluster, compute the new cluster centre for each generated cluster.

Step 4: To determine convergence, compare the newest cluster centre with the equivalent preceding cluster centre between the current and last cluster centres.

It is evident from the preceding procedure that the traditional K-means method has certain drawbacks. The biggest disadvantage is that the number of clusters, K, must be manually specified. In many circumstances, determining the proper number of clusters for a particular picture or series of photos in advance is challenging. Another disadvantage is the unpredictability with which the first clustering centre is defined. Although random selection works well when a decent set of starting clustering centres is employed, it is indeterministic and unstable, resulting in inconsistent or even failure of the picture segmentation process. There is currently no method that stops the first clustering centres from being too near to one other. To address the shortcomings of the traditional K-means technique, a method of depth picture segmentation that makes full advantage of the capabilities of Mask RCNN is suggested. To measure the accuracy of the picture segmentation and object extraction processes, proper evaluation metrics that consider the effect of image data attributes and depth image noise must be used.

The proposed algorithm's purpose is to precisely extract items of interest from a given RGB-D picture. Among the items of interest are occluded objects, in which the full object may not be visible in the image, as well as unobstructed objects.

The suggested approach contains three components:

1) An image semantic object recognition technique,

2) A K-means depth image segmentation algorithm, and 3) an object extraction tool.

In real-world visual information, numerous objects may be present in any given RGB image collected by the available sensors. These things contain both items of interest and other objects. Although segmentation can separate diverse things, it cannot discriminate between objects of interest and others. Semantic information is often utilised to recognise things of interest. To handle the challenge of identifying the value of K in the K-means algorithm, many solutions have been presented. Such systems, however, are only reliant on the utilisation of 2D pictures. The method suggested in this article, like others, takes extensive use of semantic and depth information. YOLOv3's object recognition is based on two-dimensional pictures. However, because the items in any depth image are more easily discriminated than those in a colour 2D image, depth segmentation is used in this article. Typically, there is a large difference between the impact of segmentation using the RGB picture and the effect of segmentation using the depth image. Furthermore, unlike typical image segmentation algorithms based on three colour channels, the suggested technique requires just one depth channel information to be processed, resulting in depth picture segmentation that is at least three times quicker than RGB image

segmentation. Because of the suggested method's global picture segmentation, only one operation is required to provide successful object segmentation. As a result, the suggested technique enhances overall efficiency greatly. The parallel process of picture semantic detection and segmentation is applied in this approach. To the best of our knowledge, this is the first-time semantic information has been used to address the problem of identifying the value of K in the K-means method when applied to multi-object segmentation.

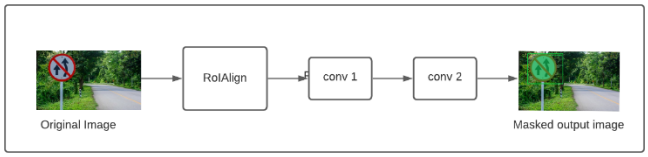


Fig 1. MASK RCNN Segmentation Framework

OBJECT EXTRACTION METHOD:

Each bounding box after picture segmentation comprises the segmentation of the related item as well as the undesirable background. The typical extraction methods discriminate between object and background using a single metric, such as size or contour. In an ideal world, each detected bounding box for each detected item has only one primary object, hence the segmentation with the single linked domain in the box is regarded as the matching object area. However, owing to picture noise and partial occlusions, there may be circumstances when the number of linked domains does not comply to the formula:

$Sc(g) = 1/cda$

The premise of Mask R-CNN is straightforward: Faster R-CNN produces two outputs for each candidate item: a class label and a bounding-box offset; we supplement this with a third branch that produces the object mask. As a result, Mask R-CNN is a natural and obvious concept. However, the additional mask output differs from the class and box outputs, necessitating the extraction of a much finer spatial arrangement of an item. Following that, we discuss the important components of Mask R-CNN, including pixel-to-pixel alignment, which is the main missing piece in Fast/Faster R-CNN.

FASTER RCNN:

To begin, we will go through the Faster R-CNN detector quickly. Faster R-CNN is divided into two steps. The first step, known as a Region Proposal Network (RPN), recommends prospective object bounding boxes. The second step, which is essentially Fast R-CNN, harvests feature from each candidate box using RoIPool and conducts classification and bounding-box regression. For speedier inference, the characteristics utilised by both phases might be pooled. A mask encodes the spatial arrangement of an input item. In contrast to class labels or box offsets, which are often compacted into short output vectors by fully connected layers, retrieving the spatial structure of masks may be handled naturally by the pixel-to-pixel correlation given by convolutions.

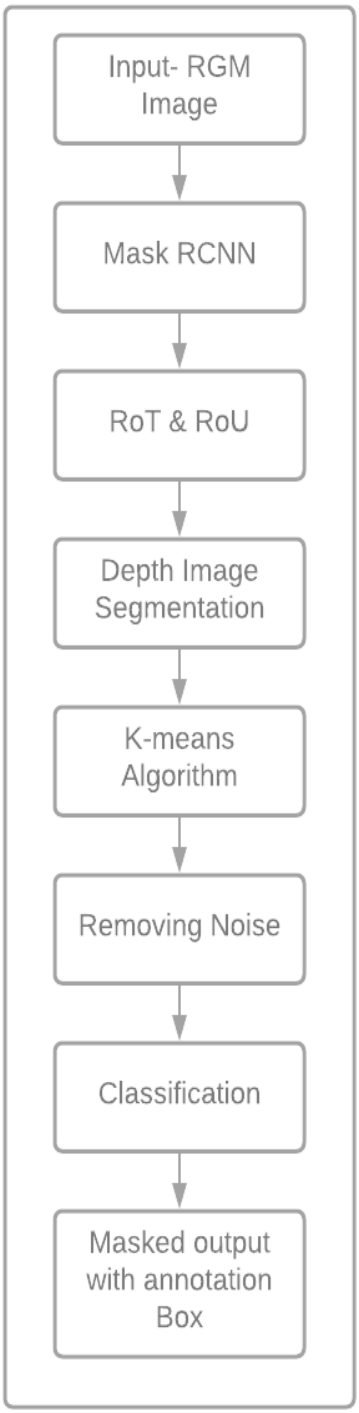


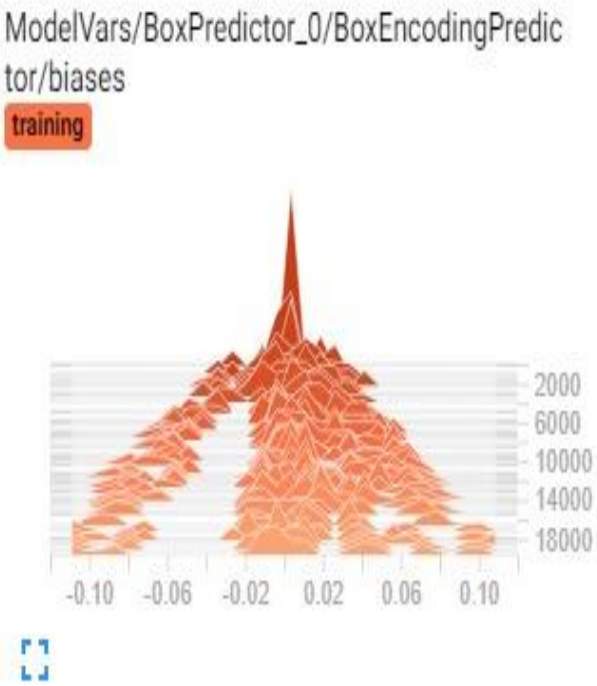
Fig 2. Schematic Representation of the proposed method

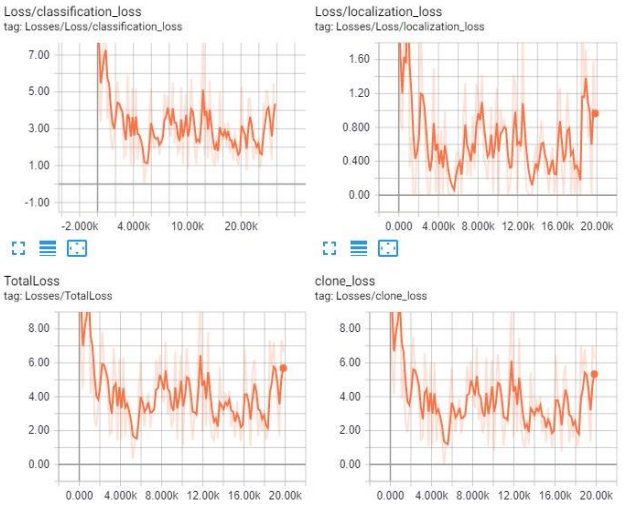
Specifically, we use an FCN to predict a m x m mask from each RoI. This enables each layer in the mask branch to keep the explicit m x m object spatial layout without compressing it into a vector representation with no spatial dimensions. Unlike prior techniques that used fc layers to predict masks, our fully convolutional representation requires fewer parameters and is more accurate, as proved by testing. This pixel-to-pixel behaviour necessitates that us RoI features, which are tiny feature maps in and of themselves, be perfectly aligned to correctly retain the explicit per-pixel spatial relationship. This inspired us to create the RoIAlign layer, which plays an important role in mask prediction. RoIPool is a common technique that extracts a tiny feature map from each RoI.

IV. RESULT & ANALYSIS

The images below demonstrate the whole picture background segmentation process, from the first input image to the final foreground item segmented using the modified K-means approach proposed in this paper. The result is disguised with an annotation box, and the precision gained with this method is astounding (greater than 85 %). We utilised 69 distinct signs,

each with 100 photos, for a total of 6,900 images in the dataset. This picture has been broken into two parts: testing and training. The training dataset contains 90% of the photos, whereas the testing dataset has 10% of the images.





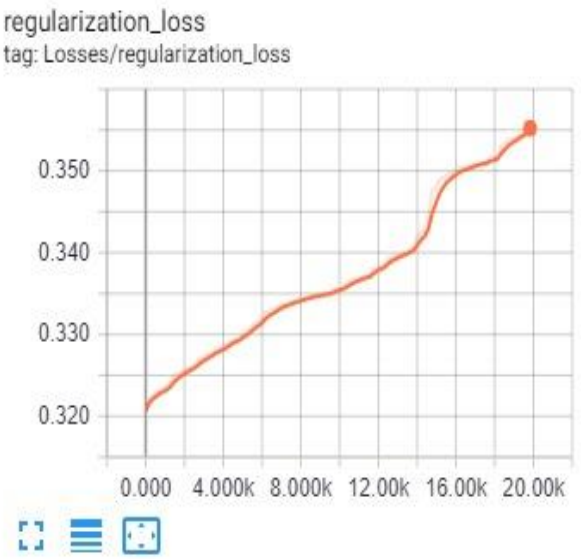
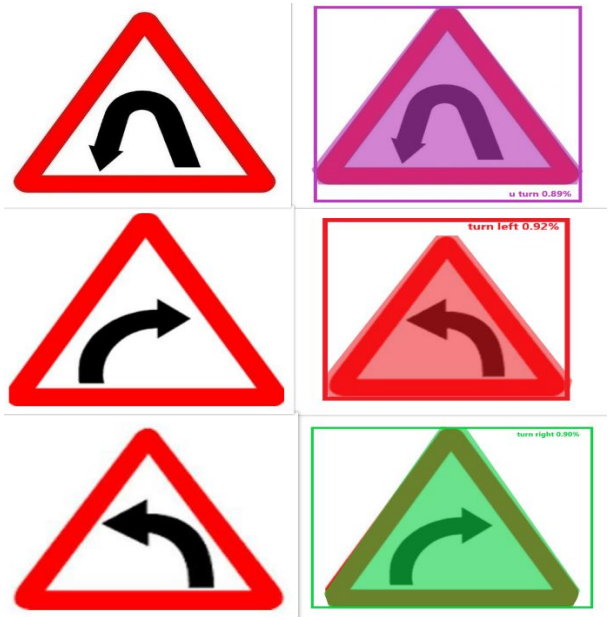


Fig 3. Learning, Loss, Encoder Graphs



Original Images Output

Fig 4. Output image 1

Fig 5. Output image 2

Original Images Output

Fig 6. Output image 3

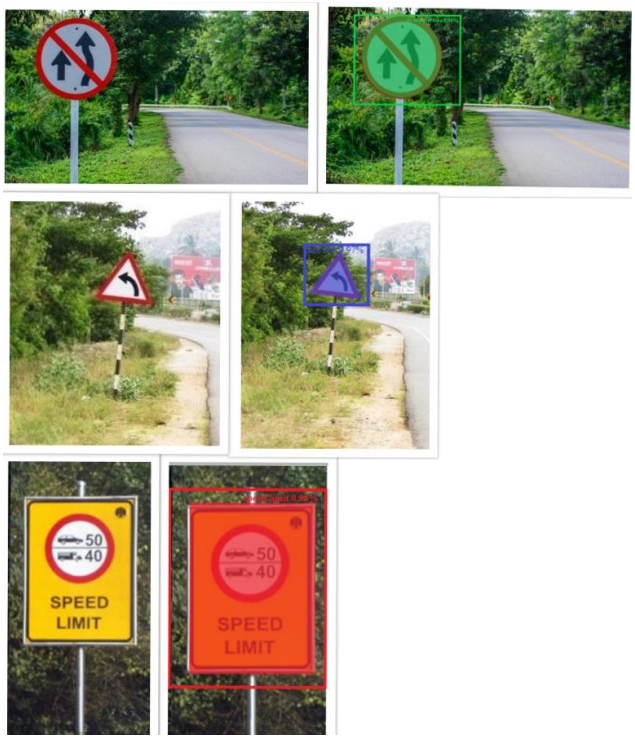


Fig 7. Output Image 3

V. CONCLUSION

An efficient picture object extraction approach is proposed in this paper. The technique includes semantic object recognition using Mask RCNN and object extraction with picture segmentation using an enhanced K-means algorithm. The K value is determined based on semantic and depth information, allowing the K-means algorithm to identify the right number of segmentations based on the actual picture. This method improves the applicability of picture segmentation in real-world scenarios. Meanwhile, using the maximin approach to identify the initial cluster centre enhances the determinacy and speed of the picture segmentation process. Future study will focus on integrating contour detection with the suggested technique to increase the accuracy of item extraction while not impacting the algorithm's processing time.

VI. REFERENCES

1) P. Dollar, R. Appel, S. Belongie, and P. Perona, ‘‘Fast feature pyramids for object detection,’’ IEEE Trans. Pattern Anal. Mach. Intell., vol. 36, no. 8, pp. 1532–1545, Aug. 2014.
2) P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, ‘‘Object detection with discriminatively trained part-based models,’’ IEEE Trans. Pattern Anal. Mach. Intell., vol. 32, no. 9, pp. 1627–1645, Sep. 2010.
3) S. Gidaris and N. Komodakis, ‘‘Object detection via a multi-region and semantic segmentation-aware CNN model,’’ in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Dec. 2015, pp. 1134–1142.
4) R. Girshick, J. Donahue, T. Darrell, and J. Malik, ‘‘Rich feature hierarchies for accurate object detection and semantic segmentation,’’ in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2014, pp. 580–587.
5) S. Ren, K. He, R. Girshick, and J. Sun, ‘‘Faster R-CNN: Towards realtime object detection with region proposal networks,’’ IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 6, pp. 1137–1149, Jun. 2017.
6) He, G. Gkioxari, P. Dollar, and R. Girshick, ‘‘Mask R-CNN,’’ in Proc. IEEE Int. Conf. Comput. Vis., Oct. 2017, pp. 2980–2988.
7) Redmon, S. Divvala, R. Girshick, and A. Farhadi, ‘‘You only look once: Unified, real-time object detection,’’ in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 779–788.
8) W. Haque, S. Arefin, A. Shihavuddin and M. Hasan, "DeepThin: A novel lightweight CNN architecture for traffic sign recognition without GPU requirements", Expert Systems with Applications, vol. 168, p. 114481, 2021.
9) S. Song, Z. Que, J. Hou, S. Du and Y. Song, "An efficient convolutional neural network for small traffic sign detection", Journal of Systems Architecture, vol. 97, pp. 269- 277, 2019. Available: 10.1016/j.sysarc.2019.01.012.
10) Vennelakanti, S. Shreya, R. Rajendran, D. Sarkar, D. Muddegowda and P. Hanagal, "Traffic Sign Detection and Recognition using a CNN Ensemble," 2019 IEEE International Conference on Consumer Electronics (ICCE), 2019, pp. 1-4
11) D. Tabernik and D. Skočaj, "Deep Learning for Large-Scale Traffic-Sign Detection and Recognition," in IEEE Transactions on Intelligent Transportation Systems, vol. 21, no. 4, pp. 1427- 1440, April 2020
12) Matoš, Z. Krpić and K. Romić, "The Speed Limit Road Signs Recognition Using Hough Transformation and Multi-Class Svm," 2019 International Conference on Systems, Signals and Image Processing (IWSSIP), 2019, pp. 89-94.
13) Degui Xiao, Liang Liu,” Super-resolution-based traffic prohibitory sign recognition ”,2019.

14) Arnab and P. H. Torr. Pixelwise instance segmentation with a dynamically instantiated network. In CVPR, 2017. 9
15) Bai and R. Urtasun. Deep watershed transform for instance segmentation. In CVPR, 2017. 9
16) S. Bell, C. L. Zitnick, K. Bala, and R. Girshick. Insideoutside net: Detecting objects in context with skip pooling and recurrent neural networks. In CVPR, 2016. 5
17) Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multiperson 2d pose estimation using part affinity fields. In CVPR, 2017. 7, 8
18) Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The Cityscapes dataset for semantic urban scene understanding. In CVPR, 2016. 9
19) Dai, K. He, and J. Sun. Convolutional feature masking for joint object and stuff segmentation. In CVPR, 2015.
20) Dai, Y. Li, K. He, and J. Sun. R-FCN: Object detection via region-based fully convolutional networks. In NIPS, 2016.