Literature Analysis on Three Dimensional Object Recognition using Deep Learning

Mrs. K. Siva Sankari
Assistant Professor, Dept. of ECE
Sri Sai Ram Institute of Technology,
Chennai, Tamil Nadu, India
sivasankari.ece@sairamit.edu.in

Dr. B. Sathya Bama
Professor, Dept. of ECE
Thiagarajar College of Engineering
Madurai, Tamil Nadu, India
sbece@tce.edu

Abstract

The subject of 3D object identification (segmentation, detection, and classification) has received much research in the areas of computer vision, graphics, and machine learning. Recently, deep learning algorithms have surpassed traditional methods for 3D segmentation problems because of their success in 2D computer vision. As a result, a number of novel methods have been developed and tested on a range of gold-standard datasets.In order for the pattern recognition system to correctly identify the item, the features must be extracted in a form that is compatible with the chosen identification technique. The location from which the features are collected does not matter. Local feature extraction and global feature extraction are the two halves of the object recognition approach.This paper provides a comprehensive analysis of the most recent advances in deep learning-based 3D object recognition. We review the most popular 3D object recognition models and evaluate their salient features.

Keywords: Deep Learning, Computer Vision, Object Recognition, 3D Objects

1. Introduction

A developing field of computer vision research is 3D object recognition. Detection of object in a given image is the primary goal of a 3D object recognition system. Many scholars have put a lot of effort into the study of 2D object identification. These days, 3D object recognition is a hot topic. 3D graphics are important in many different applications. Some examples include an intelligence surveillance system, biometric analysis, the medical field, mobile manipulation, and robots. The geometric features in 3D photographs (range images) will be more distinct than in 2D shots..

Deep learning is a complete approach that can find probable features in data without the need for manual feature engineering [1,2]. In many contexts, it is vital to be able to single out instances of objects in 3D sensory input. Three-dimensional (3D) technologies allow for more comprehensive environmental data collection. As a result, it is frequently employed in industrial detection, Augmented Reality (AR), and robotic navigation. Methods [3,4] for 3D object identification can use point clouds to learn features directly. For instance, PointNets [3,4] may either categorize the whole point cloud or predict a semantic class for each individual point. Before the development of PointNet [3,4], 3D point clouds were often flattened into 2D images or 3D voxel grids [5,6]. It is efficient in finding three-dimensional objects. However, PointNet [3,4] and PointNet++ [4] have a fault. To solve this problem, we plan on borrowing certain ideas from 2D object recognition methods, namely their attention modules.A approach based on a Gaussian Mixture Model (GMM) was put forth by Guo et al. [5] in which attention modules were enhanced by feature maps of color and others. To improve the properties of the edge information and small objects, the attention modules concentrated on fascinating places. Using an attention module built into a Region Proposal Network (RPN), as demonstrated by Fan et al. [6], the detector may zero in on specific objects with high fidelity while gathering broad contextual information with low fidelity. As a consequence of these studies, we built attention modules for use in object recognition inside 3D point clouds.

2. Related Works

The preceding 3D object identification techniques and associated attention efforts are briefly introduced in this part. We categorize our evaluations according to three different technological approaches: activation functions in neural networks, attention modules in object identification, and 3D object recognition algorithms from point clouds.

2.1. Three-Dimensional (3D) Object Detection from Point Clouds

Two-dimensional (2D) detection and three-dimensional (3D) posture estimation are both possible because to the use of three-dimensional (3D) voxel patterns (3DVPs) [7, 8]. The MV3D [9] is a multi-view 3D object detection network. Unlike the MV3Ds, Li et al. [10] and Song et al. [11] improved accuracy at the expense of a significant amount of processing.VoxelNet proposed a universal 3D detection network that is trainable from end to end and includes feature extraction and bounding box prediction[12]. With this approach, 3D object identification may be performed directly on sparse 3D points, successfully capturing 3D shape data.

2.2. Attention Module in Object Detection

Recently, certain approaches to include attention processing have been proposed to enhance CNN performance in 2D-based, massive classification applications. A Residual Attention Network was suggested by Wang et al. [13] and can integrate cutting-edge feed-forward network design. This network has the ability to continuously gather a lot of attention-related data. A Squeeze-and-Excitation module was presented by Hu et al. [14] that explicitly models channel interdependencies in order to adaptively adjust channel-wise feature responses. The computation and speed of this approach have improved. To improve accuracy, the Convolutional Block Attention Module (CBAM) [16] and the Bottleneck Attention Module (BAM) [15] introduced spatial attention. For the identification of 2D objects, these attention models worked well.

2.3. Activation Function in Neural Network

Rectified linear units (ReLUs), which have been utilized for deep networks for a while [17,18], are generally agreed to be simpler to train than logistic or tanh units. Le et al.'s observation that ReLUs seem improper for RNNs [19] was made due to the potential for huge output values to erupt out of the constrained values. Tanhhas been shown to lessen the phenomena of mean shift, according to Ang-bo et al. [20]. Li et al. [21] found that the tanh function's output may increase the values triggered by ReLU units. This motivated to use the 3D object detection network's fusion activation mechanism.

We builta unique 3D object detection network by combining the attention modules with the Frustum architecture [1,2] and the attention module [16. In order to improve accuracy, we mix the ReLUs and tanh functions in the attention module.

Model	Year	Layers	Parameters(million)
AlexNet	2012	7	62.4
VGG - 16	2014	16	6.7
GoogleNet	2014	22	6.7
ResNet-50	2015	25.6	70

Table 1 Deep Learning CNN Architectures

In KITTI detection benchmarks, the results of our suggested technique were competitive.

3. Deep Learning Architectures

Deep learning architectures serve as an important parts of the object detector. These networks take featuresout of the model's input picture. Here, we've covered a few key backbone designs seen in contemporary detectors (Table 1).Fig. 1 shows a simplified model of these CNN networks.

3.1 AlexNet

AlexNet [22], a CNN-based architecture to categorize the image is introduced by Krizhevsky et al., has won the 2012 ImageNet competition. Compared to modern models, it obtained a much greater accuracy (more than 26%). Five convolutional and three fully connected layers make up AlexNet's eight learnable layers.

3.2. VGG

Simonyan and Zisserman looked into the impact of network depth on accuracy whereas AlexNet [22] and its descendants like[23] concentrated on reduced receptive window size. They suggested VGG [24], which built networks with different depths using tiny convolution filters.

3.3. GoogLeNet/Inception

Despite advancements in categorization networks' speed and accuracy, their resource-intensive nature made it unlikely that they would be used in practical applications any time soon. The cost of processing rises exponentially when networks are scaled for improved performance. Szegedy et al. in [25] hypothesized that one of the main causes was network computation waste. Larger models tend to overfit the data since they include a lot of parameters.

3.4. ResNets

Kaiming He et al. in [26] shown how convolutional neural networks' accuracy initially reaches saturation before quickly declining. To prevent the performance from degrading, they suggested using skip connections to the stacked convolution layers. Many networks have taken inspiration from ResNets, a commonly used backbone for classification and detection.

4. Object detectors - Two stage detectors

In a two-stage detector, as shown in Fig. 1, the area suggestion process is separated into its own module. The first step of these models is to locate a collection of suggested purchases inside a picture, and the second is to categorize and localize those concepts. These systems often entail two separate processes, which results in slower idea generation, more complicated designs, and a lack of global context. Using dense sampling, one-stage detectors categorize and locate semantic items in a single pass. They use keypoints and predefined boxes of varied sizes and feature ratios to zero in on the location of objects.

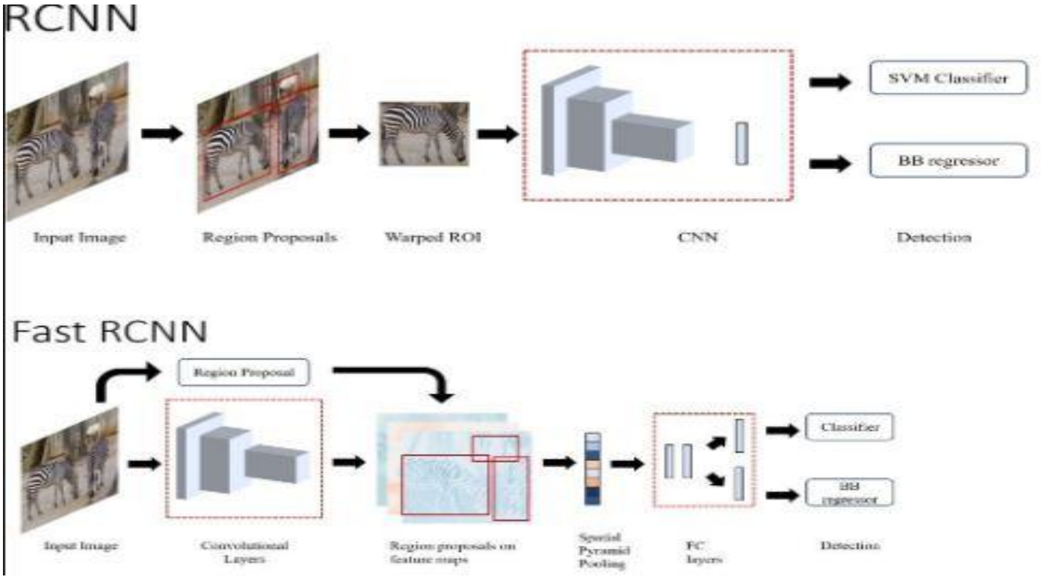

Figure 1An overview of the structure of several two-stage object detectors

.

R-CNN

Using CNNs to considerably improve detection performance was shown in the pioneering Region-based Convolutional Neural Network (R-CNN) research [27]. Using a class-independent region proposal module, R-CNN is able to convert the detection problem into a classification and localization challenge.After applying mean subtraction to an input image, this modulecreates two thousand object choices. They uses selective search to identify areas of the picture where there is a better chance of discovering an object[28]. The CNN network then warps and propagates each of these possibilities, extracting a 4096-dimension feature vector for each one.R-CNN's multi-stage training procedure is challenging. First, a big classification dataset is used to pre-train the CNN. To fine-tune, we then swap out the classification layer with a stochastically initialized N+1-way classifier using SGD, with N as the classes count[29].

Although R-CNN introduced a fresh approach to object recognition, it was sluggish (47 seconds per picture) and costly in terms of both time and resources [30]. It took days to finish the training procedure for such a small sample size, and that was with sharing some of the computations.

SPP-net

Spatial Pyramid Pooling (SPP) layer was suggested by He et al. to analyze images with any size or aspect ratio[31]. They came to understand that a fixed input was only necessary for the CNN's completely linked portion. SPP-net [32] added a pooling layer and placed CNN's convolution layers before the area proposal module. The method of selective search [33] is used to produce potential windows. The ZF-5 [34] network's convolution layers process the input image to generate feature maps.

SPP-Net has a much quicker rate of accuracy than the R-CNN model. It does not distort the subject matter due to input warping. However, it also shares some of R-CNN's shortcomings, including multistage training, high computational costs, and longer training durations, since its architecture is comparable to that of R-CNN.

Fast R-CNN

The need to train several systems independently was one of the main problems with R-CNN/SPP-Net. This was resolved by Fast R-CNN [30] by developing a single end-to-end trainable system. The network receives a picture and any suggested objects as input. The item suggestions are mapped to the picture.

quicker R-CNN

Fast R-CNN was able to recognize objects in real time, but region proposal creation was still much slower (2 seconds each picture as opposed to 0.2 seconds per image). Ren et al. [35] introduced a very complex network as a region proposal network (RPN) in [36]. There is a numerical score assigned to the likelihood of seeing an item in each of these windows. Unlike its predecessors [37], RPN introduced Anchor boxes.

R-FCN

Unlike previous two-stage detectors, which used resource-intensive approaches on each proposal, the Region-based Fully Convolutional Network (R-FCN) presented by Dai et al. [38] shared practically all computations inside the network. Convolutional layers were utilized in place of completely linked layers as an alternative. To make matters worse for localization purposes, the bottom layers of the convolutional network are not translation-invariant. The authors propose employing score maps that are position-aware to resolve the problem.

Mask R-CNN

To improve upon Faster R-CNN, Mask R-CNN [39] adds a second branch that works in parallel to do pixel-level object instance segmentation. As a completely interconnected network, the branch facilitates effective sub-pixel division when applied to the RoIs. It uses the same straightforward Faster R-CNN framework as in the classification step, but with the addition of a mask head and a bounding box regressor head for object suggestion.

Single stage detectors

YOLO was motivated by the GoogLeNet image classification model [40], and uses cascaded modules of smaller CNN[41]. Once the model has been pre-trained using ImageNet data [42] to a high degree of accuracy, it is fine-tuned by include fully connected layers and convolution with random initialization.

YOLO outperformed its single stage real time competitors by a wide margin. There were, however, several major problems with it. Its primary shortcomings were its inability to accommodate a large number of items in a single cell and its inability to accurately localize small or clustered objects. Later versions of YOLO included fixes for these problems.

SSD

To date, the only single-stage detector that can keep up with contemporary two-stage detectors like Faster R-CNN in terms of speed and accuracy is the Single Shot MultiBox Detector (SSD) [43]. VGG-16 served as the foundation for SSD, which also included extra support structures to boost speed. These supplementary convolution layers are added to the model at the very end, and their size is decreased progressively. SSD prioritizes the recognition of tiny objects early in the network when the image properties are not very simplistic, while the offsetting of default boxes and aspect ratios is handled by the more advanced layers[44].

SSD is a training method that, like Multibox[44], finds the best jaccard overlap between each ground truth box and the default box.

RetinaNet

By densely sampling the input image's features, RetinaNet [45] predicts objects. It employs two related subnets, classification and bounding box regressor, as well as the backbone of ResNet [26] enhanced by Feature Pyramid Network (FPN) [46].

5. Future trends

Over the last ten years, object detection has made significant advancements. It continues to encounter interesting challenges. Most recent object recognition models are unscalable since they need millions of bounding box annotations to train. The ability to train on data that is only weakly supervised, such as image-level tagged data, might significantly lower these costs. For autonomous driving, 3D object identification is a very important issue. Although models have attained excellent levels of accuracy, any deployment of performance below that of a person would pose safety issues.

Conclusion

While there have been tremendous improvements in object detection over the last decade, even the best detectors fall well short of their full potential. As the number of practical uses for mobile and embedded devices grows, so too will the need for lightweight models suitable for these platforms. Although there has been a rise in interest in this area, there are still many unanswered questions. We have shown how several object detectors improved over time in this article. Two-stage detectors are often more accurate, but they are too slow to be used in real-time settings like security or autonomous vehicles. However, in recent years, one stage detectors have become more quicker and similarly accurate as the former. The most accurate detector to yet is a transformer-based device called the Swin Transformer. We have great expectations for more accurate and quicker detectors given the present good trend in detector accuracy.

References

[1] Qi, C.R.; Su, H.; Mo, K.C.; Guibas, L.J. PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 77–85.

[2] Qi, C.R.; Yi, L.; Su, H.; Guibas, L.J. PointNet plus plus: Deep Hierarchical Feature Learning on Point Sets in a Metric Space. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017

[3] Qi, C.R.; Liu, W.; Wu, C.; Guibas, L.J. Pointnet: Deep learning on point sets for 3d classification and segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 652–660.

4. Qi, C.R.; Yi, L.; Su, H.; Guibas, L.J. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. arXiv 2017, arXiv:1706.02413.

5. Guo, W.; Xu, C.; Ma, S.; Xu, M. Visual attention based small object segmentation in natual images. In Proceedings of the 2010 IEEE International Conference on Image Processing, Hong Kong, China, 26–29 September 2010; pp. 1565–1568.

6. Fan, Q.; Zhuo, W.; Tang, C.K.; Tai, Y.W. Few-shot object detection with attention-RPN and multi-relation detector. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 15–20.

[7] Xiang, Y.; Choi, W.; Lin, Y.; Savarese, S. Data-driven 3d voxel patterns for object category recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1903–1911

[8] Yang, B.; Yan, J.; Lei, Z.; Li, S.Z. Aggregate channel features for multi-view face detection. In Proceedings of the IEEE International Joint Conference on Biometrics, Clearwater, FL, USA, 29 September–2 October 2014; pp. 1–8. [Google Scholar]

[9] Chen, X.; Ma, H.; Wan, J.; Li, B.; Xia, T. Multi-view 3d object detection network for autonomous driving. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1907–1915. [Google Scholar]

[10] Li, B. 3d fully convolutional network for vehicle detection in point cloud. In Proceedings of the 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Vancouver, BC, Canada, 24–28 September 2017; pp. 1513–1518. [Google Scholar]

[11] Song, S.; Xiao, J. Deep sliding shapes for amodal 3d object detection in rgb-d images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 808–816

[12] Zhou, Y.; Tuzel, O. Voxelnet: End-to-end learning for point cloud based 3d object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4490–4499.

[13] Wang, F.; Jiang, M.; Qian, C.; Yang, S.; Li, C.; Zhang, H.; Tang, X. Residual attention network for image classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 3156–3164.

[14] Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.

[15] Park, J.; Woo, S.; Lee, J.Y.; Kweon, I.S. Bam: Bottleneck attention module. arXiv 2018, arXiv:1807.06514

[16] Woo, S.; Park, J.; Lee, J.Y.; So Kweon, I. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.

[17] Nair, V.; Hinton, G.E. Rectified linear units improve restricted boltzmann machines. In Proceedings of the International Conference on Machine Learning (ICML), Haifa, Israel, 22–24 June 2010.

1. [18] Zeiler, M.D.; Ranzato, M.; Monga, R.; Mao, M.; Yang, K.; Le, Q.V.; Hinton, G.E. On rectified linear units for speech processing. In Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, 26–31 May 2013; pp. 3517–3521.

[19] Le, Q.V.; Jaitly, N.; Hinton, G.E. A simple way to initialize recurrent networks of rectified linear units. arXiv 2015, arXiv:1504.00941. [Google Scholar]

[20] Ang-bo, J.; Wei-wei, W. Research on optimization of ReLU activation function. Trans. Microsyst. Technol. 2018, 2. Availableonline: https://en.cnki.com.cn/Article_en/CJFDTotalCGQJ201802014.htm (accessed on 11 February 2021).

[21] Li, X.; Hu, Z.; Huang, X. Combine Relu with Tanh. In Proceedings of the 2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), Chongqing, China, 12–14 June 2020; pp. 51–55

[22]F. Pereira, C.J.C. Burges, L. Bottou, K.Q. W einberger (Eds.), Advances in Neural Information Processing Systems, Curran Associates, Inc. (2012), p. 9

[23]D. Fleet, T. Pajdla, B. Schiele, T. Tuytelaars (Eds.), Computer Vision – ECCV 2014, Lecture Notes in Computer Science, Springer International Publishing (2014), pp. 818-833,

[24] Muhammad U, Wang W, Chattha SP, Ali S. Pre-trained VGGNet architecture for remote-sensing image scene classification. In2018 24th International Conference on Pattern Recognition (ICPR) 2018 Aug 20 (pp. 1622-1627). IEEE.

[25]Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. and Rabinovich, A., 2015. Going deeper with convolutions.In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1-9).

[26] He, K., Zhang, X., Ren, S. and Sun, J., 2016. Deep residual learning for image recognition.In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

[27]Girshick, R., Donahue, J., Darrell, T. and Malik, J., 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 580-587).

[28]Uijlings, J.R., Van De Sande, K.E., Gevers, T. and Smeulders, A.W., 2013. Selective search for object recognition. International journal of computer vision, 104, pp.154-171.

[29] LeCun, Y., Boser, B., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W. and Jackel, L.D., 1989. Backpropagation applied to handwritten zip code recognition. Neural computation, 1(4), pp.541-551.

[30] Girshick, R., Fast R-CNN Object detection with Caffe. Microsoft Research.

[31] Grauman, K. and Darrell, T., 2005, October. The pyramid match kernel: Discriminative classification with sets of image features. In Tenth IEEE International Conference on

Computer Vision (ICCV'05) Volume 1 (Vol. 2, pp. 1458-1465).IEEE.

[32] He, K., Zhang, X., Ren, S. and Sun, J., 2015. Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE transactions on pattern analysis and machine intelligence, 37(9), pp.1904-1916.

[33] Uijlings, J.R., Van De Sande, K.E., Gevers, T. and Smeulders, A.W., 2013. Selective search for object recognition. International journal of computer vision, 104, pp.154-171.

[34]M.D. Zeiler, R. FergusVisualizing and understanding convolutional networksD. Fleet, T. Pajdla, B. Schiele, T. Tuyte laars (Eds.), Computer Vision – ECCV 2014, Lecture Notes in Computer Science, Springer International Publishing (2014), pp. 818-833,

[35] Long, J., Shelhamer, E. and Darrell, T., 2015. Fully convolutional networks for semantic segmentation.In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3431-3440).

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

[37]K. He, X. Zhang, S. Ren, J. SunDeep residual learning for image recognition2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016), pp. 770-778, 10.1109/CVPR.2016.90

[38] Dai, J., Li, Y., He, K. and Sun, J., 2016. R-fcn: Object detection via region-based fully convolutional networks. Advances in neural information processing systems, 29.

[39] He, H., Xu, H., Zhang, Y., Gao, K., Li, H., Ma, L. and Li, J., 2022. Mask R-CNN based automated identification and extraction of oil well sites. International Journal of Applied Earth Observation and Geoinformation, 112, p.102875.

[40] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. and Rabinovich, A., 2015. Going deeper with convolutions.In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1-9).

[41] Lin, M., Chen, Q. and Yan, S., 2013. Network in network. arXiv preprint arXiv:1312.4400.

[42]J. Deng, W. Dong, R. Socher, L. Li, Kai Li, Li Fei-FeiImageNet: a large-scale hierarchical image database2009 IEEE Conference on Computer Vision and Pattern Recognition, 1063-6919 (2009), pp. 248-55, 10.1109/CVPR.2009.5206848

[43]W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, A.C. BergSSD: single shot MultiBox detectorB. Leibe, J. Matas, N. Sebe, M. Welling (Eds.), Computer Vision – ECCV 2016, Springer International Publishing (2016), pp. 21-37

[44]Erhan, D., Szegedy, C., Toshev, A. and Anguelov, D., 2014. Scalable object detection using deep neural networks.In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2147-2154).

[45] Lin, T.Y., Goyal, P., Girshick, R., He, K. and Dollár, P., 2017. Focal loss for dense object detection.In *Proceedings of the IEEE international conference on computer vision* (pp. 2980-2988).

[46] Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B. and Belongie, S., 2017. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2117-2125).