CNN on Object Detection in Autonomous Vehicles: A Literature Review

Hangbo Zhang

***Abstract*—This article surveys CNN on object detection in autonomous vehicles. It listed the different areas of improvement to object detection for autonomous vehicles. Each area's pros and cons are also discussed in the article. Thus, this paper aims to become a starting point for researchers interested in using CNN on object detection for autonomous vehicles.**

***Index Terms*—convolution neural network, object detection, autonomous vehicle**

I. Introduction

In the past few years, research on autonomous vehicles has attracted lots of attention. One of the key tasks of an autonomous vehicle is object detection and tracking, which is essential for intelligent decision-making in a populated area. Accurately detecting and tracking vehicles or pedestrians on road is essential for an autonomous vehicle, such that appropriate path planning and an intelligent decision can be made, which allows an autonomous vehicle to avoid collisions and ensure the safety of passengers.[1] Object detection aims to identify the classification and location information of a given object from complex scenes; such information can then be used for complicated assignments such as subsequent tracking of the object. Moreover, in object detection, not only must object classification and positioning be simultaneously identified but also the quantity and size of objects must be determined.[2] Convolutional neural networks (CNNs) have achieved high performance in the field of object detection. However, it requires high computational power and memory which normally is limited in autonomous vehicles. Also, it does not support accurate detection at nighttime (low light conditions and lack of thermal imaging). Thus, object detection remains a challenging task in the field of autonomous vehicles.
In this paper, we explore and discuss different methods of object detection using CNN to get a better understanding and potential future avenues.

The rest of this paper is organized as follows: Section 2 describes some basic background, Section 3 describes the categorization of methods of CNN in object detection for autonomous vehicles, Section 4 describes the evaluation of the methodologies, and Section 5 describes our conclusion with some discussion.

II. Background

Convolutional neural network (CNN) is a class of artificial neural network in deep learning, most applied to analyze visual images. CNNs are regularized versions of multilayer perceptron. They take advantage of the hierarchical pattern in data and assemble patterns of increasing complexity using smaller and simpler patterns embossed in their filters. Therefore, on a scale of connectivity and complexity, CNNs are on the lower extreme.
In general, the object detection methods are grouped into two categories: two-stage detection, wherein a sparse set of object proposals are generated, and one-stage detection is a proposal-free method. Object detection models using deep learning are separated into the following two classes: regression/classification-based methods and region proposal-based methods.

Submitted on October 3, 2022. This work was intended for the ICS 691D course at the University of Hawaii at Manoa.

H. Z. Author is with the College of Natural Sciences, the University of Hawaii at Manoa, 2500 Campus Road, Honolulu 96822 USA (e-mail: hangbo@hawaii.edu).

III. CATEGORIZATION

We categorize and choose these eight papers according to the following process as shown in Figure 1.

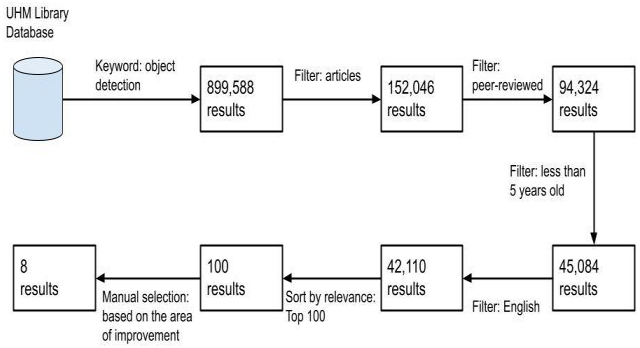

Figure 1: include & exclude process

First, we run a search on the University of Hawaii at Manoa's library database with the keyword "object detection" which returns 899,588 results. It's based on everything which includes books, patents, dissertations, articles, and conference proceedings. For the scope of this review, we exclude all the other types except articles which filter the results down to 152,046 in total. Then, to ensure the quality of the articles, we choose to only include those that are peer-reviewed, limiting the results to 94,324. As we want to review the most recent technology in the object detection area thus, we exclude all the results that are published more than 5 years ago which makes the results 45,084. Since it's a combined language database and we don't want to review the articles that are not in English which limits to 42,110 results. From those results, we manually select the top 100 which are sorted by relevance as they are greatly related to our interest. In those 100 results, we identify 8 different areas to improve object detection with the corresponding articles. The breakdown of the area is shown as follows: Training dataset: Kuznetsova [7], resolution and model weight: Chen [6], detection distance: Pang [4], detection accuracy: Wu [2], low light detection: Charan [3], harsh environments detection: Kim [1], detection speed and memory: Kim [5], and video detection: Kang [8]. The detailed result with authors, years, methodology, and area of improvement is shown in Table 1.

Table 1	**Study paper select and categorize criteria**	
Study first author, year, reference	**Methodology**	**Area of improvement**
Charan, 2020, [3]	ABiFN	low light conditions
Pang, 2019, [4]	R²-CNN	tiny object detection
Kim, 2021, [5]	1-bit weight and 8-bit activation	detection speed and memory usage
Wu, 2020, [2]	RGC Mask R-CNN	improve the accuracy of object detection
Chen, 2020, [6]	accurate yet compact deep saliency detection network	improve the resolution and reduce model weight
Kim, 2022, [1]	Robust object detection in a harsh environment	encounter harsh environment
Kuznetsova, 2020, [7]	Dataset evaluation	training dataset for models
Kang, 2018, [8]	T-CNN	extend to the video domain

Thermal image processing

Charan et al. [3] have proposed an attention-based bi-modal fusion network (ABiFN) for object detection in the thermal domain. As most of the object detection algorithms are designed to work on RGB images captured in the daytime by visible cameras, the detection performance drops with challenging lighting conditions and most of the algorithms would fail to detect in darkness as the structure and color features of the object change remarkably. They suggested that thermal IR sensors perform well in such conditions, as they are illumination invariant. Thus, they proposed this method to detect the object in the thermal images. The method contains Faster R-CNN as the base detector, the ResNet 101 architecture for feature extraction, the attention module to obtain images, and the mid-fusion to stack the feature vectors.

Tiny object detection

Pang et al. [4] have proposed a unified and self-reinforced convolutional neural network called remote sensing region-based convolutional neural

network (R^2-CNN), which is composed of the backbone Tine-Net, intermediate global attention block, and final classifier and detector, enabling the entire network efficient in both computation and memory consumption, robust to false positives, and strong to detect tiny objects. It will first crop large-scale images with a much smaller scale with 20% overlap, then applied the convolutional backbone structure to enable powerful feature extraction, then a classifier predicts the existence of the detection target, and a detector is followed to locate them accurately.

Low bit-based CNN for object detection
Kim et al. [5] have developed an optimal deep learning model to detect one class by reducing the number of channels of the object detector as much as possible for systems-on-chip (SoC) based applications. Their proposed network constructs 1-bit weights and 8-bit activations where they use the sign function to binarize the weight filter. Their proposed method for training a CNN uses 1-bit weights and 8-bit activation with two steps: forward and backward propagations. In the forward propagation, they quantize only the activations and weights except the first and last layers to reduce the accuracy loss. In the backward propagation, they simply train using the straight-through estimator (STE).

RGC Mask R-CNN
Wu et al. [2] have proposed an object detection algorithm, ResNet Group Cascade (RGC) Mask R-CNN utilizes the Pytorch framework. The characteristics of the algorithm are as follows: ResNet is adopted as the backbone, Group normalization (GN) of 32 is added to the backbone, and Cascade training is adopted. The algorithm is intended to improve Mask R-CNN to increase the accuracy of the bounding box and Mask.

Accurate yet compact deep network
Chen et al. [6] have designed an accurate yet compact deep saliency detection network based on residual learning, which achieved the smallest model size until now and makes it a better choice to be applied in subsequent applications. The proposed network includes three main components, initial saliency prediction, side-output residual learning, and top-down reverse attention.

Robust object detection in a harsh environment
Kim et al. [1] have proposed a system for solving object detection problems under a harsh autonomous-driving environment. It has two phases: first is the adversarial defense module (ADM), which is composed of an adversarial defense mechanism and reduces the computation cost and provides a robust multi-scale feature extraction in harsh environments, second is object detection with two-way DenseNet based on the feature obtained from the first phase.

Dataset evaluation
Kuznetsova et al. [7] present a large-scale dataset of images, Open Image V4, a collection of 9.2 million images annotated with unified ground -truth for image classification, object detection, and visual relationship detection. They explained how the data was collected and annotated, evaluated its quality, and reported the performance of several modern models for image classification and object detection including CNN.

Video object detection
Kang et al. [8] have proposed a deep learning framework that extends popular still-image detection frameworks to solve the problem of general object detection in videos by incorporating temporal and contextual information from tubelets. The proposed framework consists of four main components, still-image detection, multi-context suppression and motion-guided propagation, temporal tubelet re-scoring, and model combination.

IV. Evaluation

In this section, we evaluate each study for its purpose with pros and cons.

Thermal image processing
The study is trying to use thermal IR sensors to capture the thermal images while the light condition is bad due to the time of the day or weather or light switching. Then the fusion of RGB and thermal images will complement each other’s features for robust detection which has a significant improvement in the mean average precision (mAP).

Pros: it helps significantly with object detection during autonomous driving at nighttime to avoid collisions. **Cons:** thermal IR sensors will cost more energy. The image resolution of thermal images that is far away will be bad and hard to detect. The decision-making time will be very short. Stacked objects will be hard to be separated. The AP is still relatively low.

Tiny object detection
The study is using R^2-CNN to detect a tiny object from a large-scale image. **Pros**: It is efficient as it has less inference time and still preserves powerful features. It is robust as it can detect tiny objects with fewer false positives thus enabling autonomous vehicles to "see" further correctly. **Cons:** training data relies on the correct annotation and annotating all those terrible conditions very is not always possible. Even though the processing time is less compared to other models (29s), it's still significantly a lot of time for autonomous vehicles. Tracking the far object will be a big problem.

Low bit-based CNN for object detection
The study is using 1-bit weights and 8-bit activations to reduce the memory size and increase the frame per second (FPS). **Pros:** it reduces the memory size needed in autonomous vehicles thus decreasing the processing time. The high FPS will also decrease the processing time which can lead to a short decision-making time. **Cons:** it reduces the AP which means less precision with the detection.

RGC Mask R-CNN
The study is using RGC Mask R-CNN to improve the performance of Mask R-CNN. **Pros:** This method significantly increases the accuracy of the bounding box and instance segmentation which means it can detect objects more precisely. **Cons:** the performance is better with a smaller dataset but the loss of the model is greater which means it might not be well suited to autonomous vehicles.

Accurate yet compact deep network
The study is using the accurate yet compact deep saliency detection network to improve the resolution and object boundaries **Pros:** it generates sharp boundaries and coherent details of the object which means more accurate detection of the object and its actions. And it has a faster running speed compared to others. **Cons:** it's not suitable for small salient object detection which means it can be used for the nearsighted view of autonomous vehicles. It still has a large redundancy which will slow down the speed. Even though it was faster, it still needs large of time.

Robust object detection in a harsh environment
The study is using robust object detection system including ADM and two-way DenseNet to overcome the harsh environment. **Pros:** it has significant performance and speed for a harsh environment such as noise from the camera, driving state, weather, and system perspective. **Cons:** While they applied AT, both accuracy and speed will decrease. The model will only learn the situations defined in advance, so it can not deal with unexpected situations which are very common in autonomous driving.

Dataset evaluation
The study evaluates the performance of two modern object detections on Open Images, the two-stage Faster-RCNN, and the single-shot SSD. They reported generally, the performance of all detectors continuously improves by adding more training data. A well-labeled and classified dataset will provide more than enough training data to reach the model performance limits. **Pros:** the model will require less training to reach the performance limit. The more high-quality training data provide, the better performance autonomous vehicles will be. **Cons:** it will require large amounts of data and labeling.

Video object detection
The study is using a deep learning framework that incorporates temporal and contextual information to detect objects in video. **Pros:** autonomous vehicles are using cameras to capture videos and then divided them into images, this method will reduce the transformation time and gain continuous object detection. Also, it will be beneficial to track the object in the live video then just images. **Cons:** it will require more memory size. There will be a lot of redundancy among video frames. Tracking and predicting the location is still a hard problem.

V. Conclusion

In this review, we have explored 8 studies with CNN and object detection in the autonomous vehicles field. They all have different areas of improvement for autonomous vehicles. We discussed their methodologies and the pros and cons of each study in the autonomous vehicles field. Further study of each would be greatly useful for autonomous driving. Some combination of the studies will also improve the performance of autonomous vehicles.

References

[1] Y. Kim, H. Hwang, J. Shin, "Robust object detection under harsh autonomous-driving environments", IET Image Process. 2022; 16:958–971.

[2] M. Wu, H. Yue, J. Wang, Y. Huang, M. L, Y. Jiang, C. Ke, C. Zeng, "Object detection based on RGC mask R-CNN", IET Image Process., 2020, Vol. 14 Iss. 8, pp. 1502-1508

[3] A. S. Charan, M. Jitesh, M. Chowdhury, H. Venkataraman, "ABiFN: Attention-based bi-modal fusion network for object detection at night time", ELECTRONICS LETTERS 26th November 2020 Vol. 56 No. 24 pp. 1309–1311

[4] J. Pang, C. Li, J. Shi, Z. Xu, H. Feng, "R^2-CNN: Fast Tiny Object Detection in Large-Scale Remote Sensing Images", IEEE Transactions on Geoscience and Remote Sensing, Vol. 57, No. 8, August 2019

[5] Y. Kim, O. Choi, W. Hwang, "Low bit-based convolutional neural network for one-class object detection", Electronics Letters, March 2021 Vol. 57 No. 6

[6] S. Chen, X. Tan, B. Wang, H. Lu, X. Hu, Y. Fu, "Reverse Attention-Based Residual Network for Salient Object Detection", IEEE Transactions on Image Processing, Vol. 29, 2020

[7] A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Malloci, A. Kolesnikov, T. Duerig, V. Ferrari "The Open Images Dataset V4: unified Image Classification, Object Detection, and Visual Relationship Detection at Scale", International Journal of Computer Vision (2020) 128:1956–1981

[8] K. Kang, H. Li, J. Yan, X. Zeng, B. Yang, T. Xiao, C. Zhang, Z. Wang, R. Wang, X. Wang, W. Ouyang, "T-CNN: Tubelets with Convolutional Neural Networks for Object Detection From Videos", IEEE Transactions on Circuits and Systems for Video Technology, Vol. 28, No. 10, October 2018