Surveillance System Scheme using Multi-detection Attribute with Optimized Neural Network Algorithm on Intelligent Transportation System

1st Akhmad Yusuf Nasirudin
School of Electrical Engineering
Telkom University
Bandung, Indonesia
akhmadyusufn@student.telkomuniversity.ac.id

2nd Korediantо Usman
School of Electrical Engineering
Telkom University
Bandung, Indonesia
korediantousman@telkomuniversity.ac.id

3rd Suryo Adhi Wibowo
School of Electrical Engineering
Telkom University
Bandung, Indonesia
suryoadhiwibowo@telkomuniversity.ac.id

***Abstract*— Intelligent Transportation System (ITS) combines a transportation system with Information and Communication Technology (ICT) system, where ICT system plays a role in adding functionality in the form of intelligence resembling human intelligence to the transportation system. The combination allows humans to know the real state of the transportation system including transportation components, such as the status of the road, objects around the vehicle, and the state of the vehicle, thus enabling humans to optimize the transportation system. For example, if there is a group of thief that using a van on the road, we can fasten the process to detect where is the route that used by the thief by adding a vehicle detector on the traffic light camera. This detector will be work better if the detector can detect the van in real-time and in a high resolution image. This work will discuss on how to increase the detector system performance on inference time (fps) and accuracy using HRNet and FCOS. HRNet is a high resolution image network architecture that can process image in a multiple resolution (low, medium, high) to maintain the high resolution but still have an enough image feature to process, while FCOS is a one stage anchor-free detector, so it can detect the object faster than the anchor-based detector. The performances was even more better when we add a warm up training before the training process. Our experimental results shows that our system has a better result compared with the reference result using same dataset and hyperparameter. It also has a better result compared with the reference result that using the reference dataset and hyperparameter.**

***Keywords*— intelligent transportation system; objet detection; vehicle detection; attribute detection; computer vision; image processing; surveillance system.**

I. PRELIMINARY

Nowadays, the rapid growth of vehicles is a common problem in Indonesia. This situation is affected by people's lifestyles, especially in urban areas, where people need a fast movement from one place to another. It causes inevitable traffic problems such as traffic congestion, traffic accidents, delayed departure schedules, and greater vehicle emissions. Several solutions were offered to overcome these problems, one of them is the safety system implementations such as airbags, safety belt, and the constant improvement of road and highway construction. However, building roads is not the right solution to cut traffic congestion, because it is very expensive and requires a very large space which becomes the urban area's limitations. On the other hand, transportation infrastructure is very important for economic growth. Therefore, this work will try to figure out a compromise solution to reduce the traffic crimes. Intelligent Transportation System (ITS) is a global topic that attracts worldwide attention from transportation professionals, automotive industries, and politics. ITS combines the transportation system with information and communication technology (ICT) systems to carry out advanced communication, information, and electronic technology that can solve the traffic problems. By utilizing ITS technology, vehicles become more efficient, environmentally friendly, and certainly safer.

ITS utilizes several fields of technology to support traffic functions, one of them is surveillance system technology that utilizes object detection feature. A surveillance system is an important technology that must be implemented in traffic because it can prevent and overcome traffic crimes such as hit-and-run, robbery, theft, etc. A large number of cases and the amount of time needed to solve caused many cases to be delayed and even dismissed. Therefore, a computer vision-based traffic surveillance system is needed with expectations to help to prevent and resolve traffic crimes.

In terms of object re-identification, most previous works focus on a human face or person identification. Different from the face or person re-identification, vehicle re-identification is more challenging as it is very difficult to discriminate vehicles with a similar visual appearance which belong to the same model. In other words, the inter-class (inter-ID) difference can be very subtle. There even exist newly produced vehicles which look the same. Usually, vehicle re-identification can be done by detecting and reading

the vehicle plat number, but in any case, the plate number can be faked or not recognizable due to the image resolution.

Based on the explanation above, an object detection-based method to do the vehicle re-identification task should be proposed by analyzing the vehicle attributes like the front glasses, an object that sticks on the front glasses, etc. To do its job properly, the detection system that applied to the traffic control system must be able to work quickly and accurately This work will also modify the loss function of the method so it can increase the accuracy of the detector.

In this past few years, vehicle attribute detection and classification [1],[2] receive an increasing research interest is image processing and computer vision because of the importance of surveillance system implementation. Vehicle attribute detection and classification which locate and classify vehicles that are captured by different camera, is an important and challenging problem in intelligent transportation system.

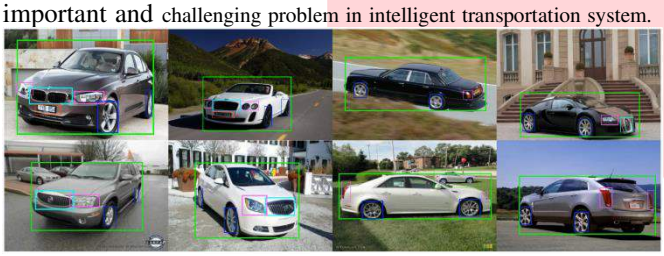

FIG 1
(Example of vehicle and its attribute being detected by system)

In terms of object attribute detection, most of previous works was focused on face or person detection in the reference [3], [4],[5]. Contrasting from face or person attribute detection, vehicle attribute detection is more challenging as it is very hard to separate vehicle with similar optical presence with belong to same model and brand. Although the license plate number contains a unique ID for a vehicle, sometimes it is not recognizable.

Inspired by this, we would like to support the vehicle attribute detection to increase its performance. We implement our system to a dataset to find out how good is our system, and the results were not disappointing. In this research, we first collect a vehicle dataset from the subset of Open Image Dataset (OID) [6] and then we do a preprocessing step to match the dataset format with the detector and backbone that we use.

At this time, Convolutional Neural Network(CNNs)-based detection methods [7],[8],[9] have show a great performance in terms of detection performance. To improve the performance of the detector, we add a "warm up training" to reduce the early-over-fitting and reducing the learning rate to reduce the under-fitting that might be occur to the model. Our experimental results show the improvement of the detection performance with this modification.

II. METHODS AND MATERIAL

A. *Related Works*

With the development of tools that allow for neural computing, many researchers wish to design object detection using deep learning methods. At present, most advanced object detection is mostly anchor-based [9]. Faster RCNN is a CNN-based detector derived from RCNN[10] and Fast RCNN [11]. All of these ongoing studies follow a similar two-step procedure, finding the rough areas where objects are most likely to be found, then rebuilding these boxes and classifying the objects. However, Faster RCNN implements the procedure only with convolutional neural networks, increasing accuracy and reducing inference time significantly. While SSD prioritizes increasing detection speed by integrating two steps into one. Apart from the time efficiency of SSD's, for traffic surveillance scenarios, vehicles on the road are relatively small in size and may be covered by other vehicles. There is still a lot of room to improve performance by overcoming these challenges.

Several results have been reported in the literature on the tasks of vehicle detection, vehicle plate detection, and vehicle logo detection. Traditional vehicle detection methods usually use information about geometric features [12], symmetry [13], [14], texture [15], and color [16]. DCNN has also shown excellent results in classification and detection. Fans et al. implemented Faster R-CNN a method for detecting generic objects, to improve the performance of Faster R-CNN on vehicle detection and achieve excellent results on the KITTI vehicle dataset.

As for vehicle plate detection, Wu and Li[17] proposed a way to detect license plates by selecting potential frames. Chang et al.[18] detects license plates based on number plate geometry features, namely shape, symmetry, height-to-width ratio, color and texture. Gerber and Chung et al.[19] designed a way for multi-CNN to detect license plates using mobile devices. A CNN will be used to verify the car and then the output of that CNN will be given to the next supervised CNN for vehicle plate detection.

For vehicle attribute detection, most of the literature focuses on extracting strong discriminatory features, such as brand, color, shape, or vehicle type, and even information that is combined to maximize attribute detection results[20],[21]. Several existing studies also assign different weights to different features which serve to better differentiate each vehicle [22],[23].

As already explained, all detection methods on vehicles are handled separately. Among the collection of object detection methods, those based on deep learning have shown very good results, namely FCOS [24], SSD, Mask R-CNN, Faster R-CNN, R-FCN[25], and YOLO[26].Where, the network architecture can be divided into one-stage and two-stage region-proposal based. Among these methods, FCOS is one of the methods that have a pretty good accuracy and also very efficient processing time.

Since attribute detection is a procedure whose results can be used for further tasks, in this work, we chose FCOS as the basis of our detection for the reasons previously mentioned. Furthermore, since we are working on high-quality images, you will be using HRNet[27] for the image extraction architecture. We also used focal loss[28] to increase mAP.

B. *Dataset*

In this research, we use two datasets, namely Open Image Dataset (OID) and Car Parts Dataset [29]. The OID is a fast dataset owned by google. This dataset has more than 59

million images with more than 15 million bounding boxes divided into 600 classes. We can use this dataset to perform

FIG 4
(Sample of OID used in this research.)

detection, segmentation, and relationships. For this research, we will only use 30 thousand images which are divided into 8 classes, namely 'bicycle', 'car', 'motorcycle', 'airplane', 'bus', 'train, 'truck', 'boat'. Fig.4 is the sample of OID dataset classes that used in this research. We also do some pre-processing for this dataset such as:

1. The dataset was divided into train, validation, and test.
2. There are two processes of dataset division. First, dataset will be divided into train and test. Second, from the train dataset, it will be divided again into train and validation.
3. The dataset annotation was converted into COCO format to fit the detector input.

Car parts dataset is a dataset containing labels, bounding boxes, and areas of car parts. This dataset was used in[29] to test its performance on semantic segmentation using several deep learning methods such as Mask R-CNN, HTC[30], CBNet[31], PANet[32], and GCNet[33]. This dataset consists of 500 images of cars (sedans, trucks, and SUVs) which are divided into 18 classes, namely 'back bumper', 'back glass', 'back left door', 'back left light', 'back right door', 'back right light', 'front bumper', 'front glass', 'front left door', 'front left light', 'front right door', 'front right light', 'hood', 'left mirror', 'right mirror', 'tailgate', 'trunk' and 'wheel'.

We can see in the Fig.5, that the dataset only show the segmentation area, but actually in the annotation, there is also the labels and bounding box. This dataset annotation uses the COCO Dataset format where we can perform detection and segmentation using this dataset. There are three points of view in this dataset, front, side, and back.

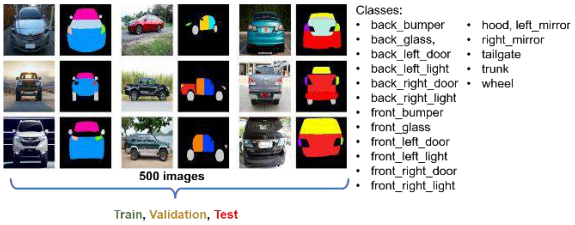

FIG 5
(Sample of Car Parts Dataset used in this research.)

C. *Vehicle and Attribute Detection*

Most of state-of-the-art extraction network convert image into medium to low-resolution images, whereas high-resolution imaging is important in vision problems, such as human pose estimation, semantic segmentation, and object detection. Previous work encodes the input image into low-resolution imaging via a subnetwork formed by linking high-to-low resolution convolutions in a series (e.g., VGGNet, ResNet), then returns high-resolution imaging from the encoded low-resolution imaging. The High-Resolution

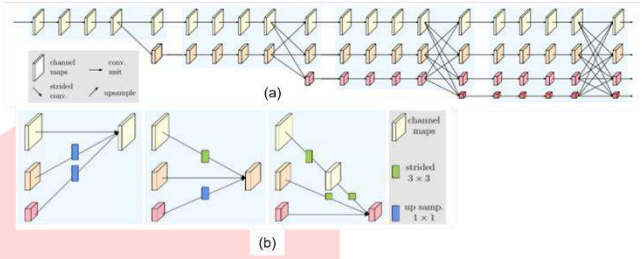

FIG 6
(The illustration of HRNet on multi-level resolution feature extraction)

Network (HRNet) works differently in restoring high-resolution imaging. Instead of linking high-to-low resolution in series, HRNet links high-to-low resolution in parallel. HRNet also performs an iterative exchange of information between resolutions \cite{wang2020deep}. The advantage of using this method is that the information in the feature becomes richer. Fig 4 (a) explains the example of HRNet. There are four stages. The 1st stage contains of high-resolution convolutions, while the 2nd (3rd, 4th) contains of the two (three , four)-resolution convolution blocks. Meanwhile the Fig. 6(b) is the illustration of the fusion of information for high, medium, and low resolution. Right legend: strided 3 × 3 = two-strided 3 × 3 conv., up samp. 1 × 1 = bilinear up sampling followed by 1 × 1 conv.

The most popular anchor-free detector today is YOLOv1. Instead of using anchor boxes, YOLOv1 predicts the location of the bounding box at a point close to the object's center because points close to the object's center are considered to produce better quality detection. However, because only points close to the center of the object are used to predict the bounding box, YOLOv1 has low recall as described in YOLOv2[34] . In the end, YOLOv2 also uses anchor boxes. Learning from YOLOv1, FCOS takes all points in the ground truth bounding box to predict the bounding box, while the low-quality bounding box is handled using the center-ness branch. As a result, FCOS has good performance compared to anchor-based detectors.

With the proposed dataset, we perform the vehicle detection using an anchor-free one stage detector Fully Convolutional One-Stage Object Detection (FCOS). The difference between anchor-based and anchor-free object

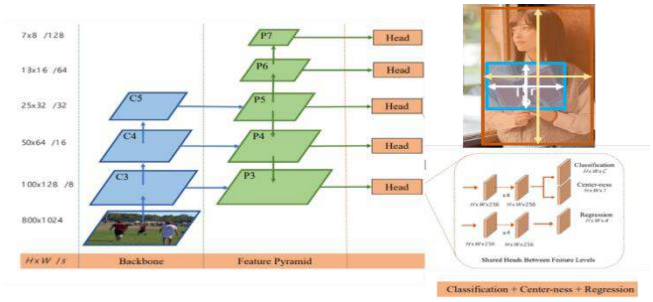

FIG 7
(The illustration of FCOS object detection.)

detector is anchor-based object detection will generate a lot of bounding box and find the best IoU value[35], so there will be many redundant computation. Meanwhile anchor-free object detection will find the center of the object and try to regress the bounding box value. As shown in the Figure 7, FCOS works by making a prediction on 4D vector (l, t, r, b) encoding the location of the bounding box at each object pixel by learning from supervised dataset with ground-truth bounding box information during training. The equation to calculate the centerness is as follows:

$$centerness = \sqrt{\frac{\min(l,r)}{\max(l,r)} \times \frac{\min(t,b)}{\max(t,b)}} \qquad (1)$$

where l is the distance from center to bounding box left side, r for right side, t for top side, and b for bottom side.
Fig. 5 also shows the network architecture of FCOS, where C3, C4, and C5 is the feature maps genereated by the backbone network (HRNet) and P3, P4, P5, P6, and P7 are the level of feature used for the final prediction. H × W is the height and width of feature maps, /s (s =8, 16, ...,128) is the downsampling ratio of the feature maps in the input image.

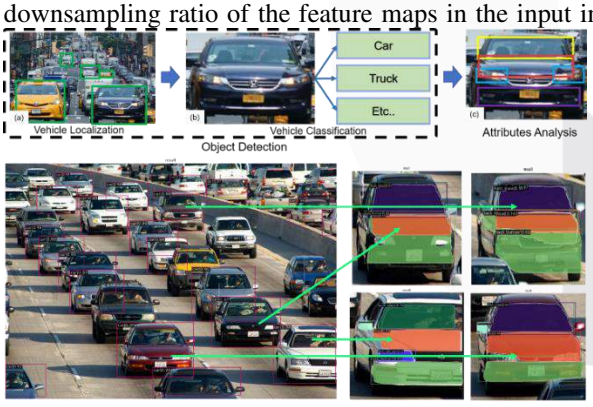

FIG 8
(Detection scheme conducted in this research)

Fig. 8 is the illustration of the detection scheme used in this research, the steps are as follows:

1. The detector is doing a localization on an image contain of some vehicle objects.
2. The classification process of every detected object.
3. After the object was classified and localized, the surveillance system will perform an attribute analysis to recognize the vehicle specifically.

Warm up training is a pre-training process using a low-value learning rate before the data trained using the regular learning rate. It helps the network to slowly adapt to the data intuitively and reduce the primacy effect of the early training examples. If the data set is highly differentiated, it can suffer from a sort of early over-fitting. If the shuffled data happens to include a cluster of related, strongly-featured observations, the model's initial training can skew badly toward those features or worse, toward incidental features that aren't truly related to the topic at all.

III. RESULT AND DISCUSSION

In this section, we conduct experiments to demonstrate the performance of the methods that have been carried out which consist of vehicle detection and attribute detection.

We carried out a series of experiments using the Open Image Dataset for our vehicle detection task. The image resolution used in this experiment is 1300$\times$800. The dataset was divided into train, validation, and test. There are two processes of dataset division. First, the dataset will be divided into train and test. Second, from the train dataset, it will be divided again into train and validation. This experiment was carried out using a torch-based object detection tool called MMDetection with an HRNet backbone network. We also use a pre-trained model that has been trained using the COCO dataset with the HRNet backbone architecture and FCOS as the detector, this is because using a pre-trained model the desired results will be easier to achieve.

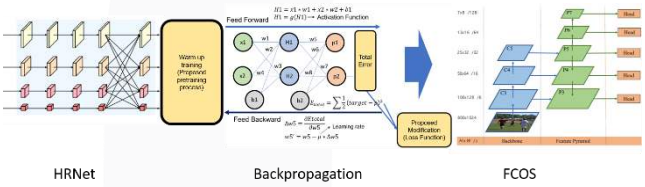

FIG 9
(The proposed training flow on the model.)

Fig.9 explains that after the system do the feature extraction in HRNet, the feature will be used to train the model using backpropagation method. But before the backpropagation process, we add a warm up training to prevent overfitting when the system train the model.

Experiment Detail: We experimented by modifying FCOS+OID by using warm-up training. We added a "warm-up training" to be able to reduce the early-over-fitting and under-fitting on the model.

The table 1 explains that the experiments training was monitored every 3 epochs (3, 6, 9, 12, 15). The number of images used is 30,000 images divided into three parts train, test, and validation with a comparison of 80% percent train, 20% test, and validation is taken from 20% of the train. The results with the best performance were obtained with warm up training using warm up iteration=1000 and warm up ratio=0.001 with the result mAP=51.5 and inference time=5.6.

After the experimental results were obtained, we took the configuration that had the best performance and then used the

configuration on FCOS+OID without modification, Faster-RCNN+OID without modification, and Faster-RCNN+OID with modification (focal loss and warm-up training) and compared with our model. From the comparison results, our model has the highest mAP and also the best inference time as shown at Table II.

TABLE II
(mAP and Inference Time Comparison Result)

Method	Backbone	mAP	Inf.Time (fps)
FCOS+OID	HRnet	50.8	5.8
Faster RCNN+OID	HRnet	50.5	4.1
mod FCOS+OID	HRnet	51.	5.6
mod Faster RCNN+OID	HRnet	52.1	4.2

The attribute detection configuration uses the best configuration of vehicle detection with parameters gamma=2, alpha=0.25, and epoch=12 but uses the Car Parts dataset. Here is the result example of attribute detection.

FIG 10
(The example of attribute detection result.)

This work can be a complement to the previous work [36], because the previous work also performed an object detection only using a different method. With the method that the author uses, of course the previous work can be a more thorough object detection.

FIG 11
(Result image from the previous works)

Figure 11 is the result of previous work, in which the detection has not yet reached the pixel stage. With the author's method, the detection can be deeper so that the system can detect it more accurately and the detection results will be better.

IV. CONCLUSIONS

This research proposed HRNet architecture to maintain the feature maps resolution, so it can detect even a small object in the image.

The detector used in this research is FCOS detector, a one-stage and anchor-free detector to keep the fps high so the surveillance system can process a moving object in a great way. For comparison, Faster R-CNN detector as a two-stage detector has been used. The dataset is contain about 30.000 images that divided into train, validation, and test taken from Open Image Dataset.

By adding a “training warm up”, reducing the learning rate, and replacing the loss function with a focal loss, the model mAP can be improved with only sacrificing a very small value of inference time. When the detector has a good mAP and a high fps, the detector can do a detection well on a small moving object, like in the traffic area.

REFERENCE

[1] Fang, Wenhua, et al. "Vehicle re-identification collaborating visual and temporal-spatial network." *Proceedings of the fifth international conference on internet multimedia computing and service*. 2013..

[2] Zapletal, Dominik, and Adam Herout. "Vehicle re-identification for automatic video traffic surveillance." *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 2016.

[3] Schroff, Florian, Dmitry Kalenichenko, and James Philbin. "Facenet: A unified embedding for face recognition and clustering." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015..

[4] Ding, Shengyong, et al. "Deep feature learning with relative distance comparison for person re-identification." *Pattern Recognition* 48.10 (2015): 2993-3003.

[5] Lin, Weiyao, et al. "Learning correspondence structures for person re-identification." *IEEE Transactions on Image Processing* 26.5 (2017): 2438-2453.

[6] Kuznetsova, Alina, et al. "The open images dataset v4." *International Journal of Computer Vision* 128.7 (2020): 1956-1981.

[7] Liu, Wei, et al. "Ssd: Single shot multibox detector." *European conference on computer vision*. Springer, Cham, 2016.

[8] He, Kaiming, et al. "Mask r-cnn." *Proceedings of the IEEE international conference on computer vision*. 2017.

[9] Ren, Shaoqing, et al. "Faster r-cnn: Towards real-time object detection with region proposal networks." *Advances in neural information processing systems* 28 (2015)..

[10] Girshick, Ross, et al. "Rich feature hierarchies for accurate object detection and semantic

segmentation." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014.

[11] Girshick, Ross. "Fast r-cnn." *Proceedings of the IEEE international conference on computer vision*. 2015.

[12] Bertozzi, Massimo, Alberto Broggi, and Stefano Castelluccio. "A real-time oriented system for vehicle detection." *Journal of Systems Architecture* 43.1-5 (1997): 317-325..

[13] Gao, Yongbin, and Hyo Jong Lee. "Vehicle make recognition based on convolutional neural network." *2015 2nd International Conference on Information Science and Security (ICISS)*. IEEE, 2015./

[14] Kalinke, Thomas, Christos Tzomakas, and Wemer von Seelen. "A texture-based object detection and an adaptive model-based classification." *Procs. IEEE Intelligent Vehicles Symposium*. Vol. 98. Citeseer, 1998.

[15] Buluswar, Shashi D., and Bruce A. Draper. "Color machine vision for autonomous vehicles." *Engineering Applications of Artificial Intelligence* 11.2 (1998): 245-256.

[16] Fan, Quanfu, Lisa Brown, and John Smith. "A closer look at Faster R-CNN for vehicle detection." *2016 IEEE intelligent vehicles symposium (IV)*. IEEE, 2016..

[17] Ahmed, Mohammed Jameel, et al. "License plate recognition system." *10th IEEE International Conference on Electronics, Circuits and Systems, 2003. ICECS 2003. Proceedings of the 2003*. Vol. 2. IEEE, 2003.

[18] Chang, Shyang-Lih, et al. "Automatic license plate recognition." *IEEE transactions on intelligent transportation systems* 5.1 (2004): 42-53.

[19] Gerber, Christian, and Mokdong Chung. "Number plate detection with a multi-convolutional neural network approach with optical character recognition for mobile devices." *Journal of Information Processing Systems* 12.1 (2016): 100-10.

[20] Liu, Hongye, et al. "Deep relative distance learning: Tell the difference between similar vehicles." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.

[21] Liu, Xiaobin, et al. "Ram: a region-aware deep model for vehicle re-identification." *2018 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2018.

[22] Wang, Zhongdao, et al. "Orientation invariant feature embedding and spatial temporal regularization for vehicle re-identification." *Proceedings of the IEEE international conference on computer vision*. 2017.

[23] Zhou, Yi, Li Liu, and Ling Shao. "Vehicle re-identification by deep hidden multi-view inference." *IEEE Transactions on Image Processing* 27.7 (2018): 3275-3287.

[24] Tian, Zhi, et al. "Fcos: Fully convolutional one-stage object detection." *Proceedings of the IEEE/CVF international conference on computer vision*. 2019.

[25] Dai, Jifeng, et al. "R-fcn: Object detection via region-based fully convolutional networks." *Advances in neural information processing systems* 29 (2016).

[26] Redmon, Joseph, et al. "You only look once: Unified, real-time object detection." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.

[27] Wang, Jingdong, et al. "Deep high-resolution representation learning for visual recognition." *IEEE transactions on pattern analysis and machine intelligence* 43.10 (2020): 3349-3364.

[28] Lin, Tsung-Yi, et al. "Focal loss for dense object detection." *Proceedings of the IEEE international conference on computer vision*. 2017.

[29] Pasupa, Kitsuchart, et al. "Evaluation of deep learning algorithms for semantic segmentation of car parts." *Complex & Intelligent Systems* (2021): 1-13..

[30] Chen, Kai, et al. "Hybrid task cascade for instance segmentation." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019..

[31] Liu, Yudong, et al. "Cbnet: A novel composite backbone network architecture for object detection." *Proceedings of the AAAI conference on artificial intelligence*. Vol. 34. No. 07. 2020.

[32] Wang, Kaixin, et al. "Panet: Few-shot image semantic segmentation with prototype alignment." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019.

[33] Cao, Yue, et al. "Gcnet: Non-local networks meet squeeze-excitation networks and beyond." *Proceedings of the IEEE/CVF international conference on computer vision workshops*. 2019

[34] Redmon, Joseph, and Ali Farhadi. "YOLO9000: better, faster, stronger." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.

[35] Zhang, Shifeng, et al. "Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020.

[36] Zhao, Yanzhu, et al. "Structural analysis of attributes for vehicle re-identification and retrieval." *IEEE Transactions on Intelligent Transportation Systems* 21.2 (2019): 723-734.