Article

Polygonal Dataset for Vehicle Instance Segmentation for Private Surveillance System

Najmath Ottakath [1,†,‡] and Somaya Al-Maadeed [1,*]

[1] Affiliation 1; Qatar University
* Correspondence: gonajmago@gmail.com;
‡ These authors contributed equally to this work.

Abstract: Vehicle identification is an important task in traffic monitoring because it allows for efficient inference and provides a cause for action. Vehicle classification via deep learning and other approaches such as segmentation is a critical tool for re-identification. In this paper, instance segmentation is used for vehicle make identification with license plate detection, allowing for better unique vehicle recognition for re-identification. An existing dataset is re-annotated and modified for polygonal segmentation of the vehicle's unique frontal features, resulting in representation of the vehicle with its frontal form learned. In addition, an additional license plate identification class is added for efficient re-identification further down the re-identification and tracking pipeline. Furthermore, an additional class of license plate identification is added for efficient re-identification further down the re-identification and tracking pipeline. The results showed improved classification as well as a high mAP for the dataset when compared to previous approaches based on CNN and deformed CNN. Furthermore, a deep residual network and fully connected layer-based classification were utilized as the backbone for feature representation. Instance segmentation detects objects by segmenting and classifying regions of interest. The imbalance in the dataset is resolved using a mosaic-tiled approach, which produces greater precision than other approaches evaluated for in the paper.

Keywords: Instance segmentation; Classification; Vehicle make classification; Mosaic-tiled augmen-tation

1. Introduction

Vehicle surveillance is an essential task in public security. Unique features of vehicles like vehicle make, model, and license plate are typically utilized to identify the vehicle. With traffic cameras at every junction of the streets, the entrance of high-security buildings, parking lots, and public places, there is an opportunity to surveil and track the traffic while monitoring the road. Images and/or videos are captured to provide a plethora of opportunities through scene understanding object detection, recognition, and segmentation using automated approaches such as image processing, machine learning, and deep learning techniques known as computer vision [1,3,5]. Further subtasks are performed from these approaches such as re-identification[8] , tracking, and similarity matching[4,6,7]. Transfer learning has been widely utilized on existing pretrained models for video surveillance [2]. Machine learning and deep learning models were applied to vehicle data to infer the make, model, and license plate region [13]. In each case either the wholesome image was used for analysis, or a region of interest was carved where rectangular boundaries were drawn, to identify the exact location of the contextual features to categorize or re-identify[8]. In the context of cars, the car make was defined by the front of the car in [24]. The region of interest was extracted from this dataset to identify the car's make and model. This enabled a better representation of the uniqueness of the car. Further, the license plate was also extracted which further can be fed to an ALPR (Automatic license plate recognition) system for digit recognition.

 doi:10.20944/preprints202212.0475.v1

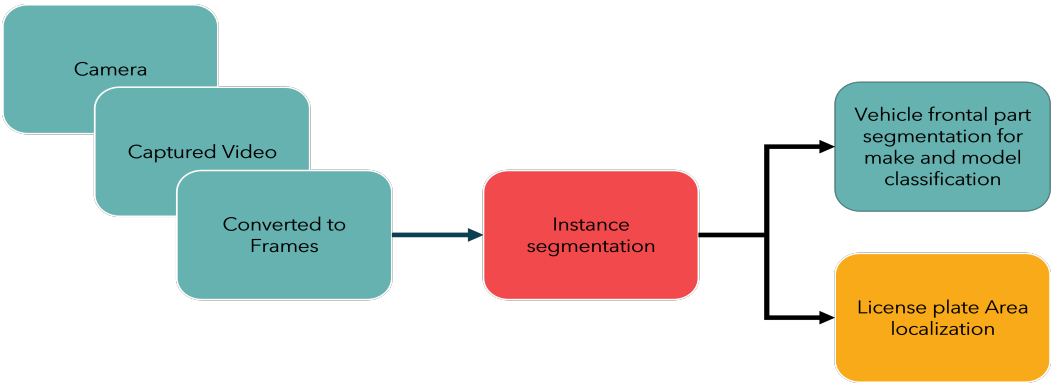

Figure 1. Vehicle instance segmentation technique for license plate localization along with make and model classification.

In computer vision, the region of interest extraction has been widely performed by segmentation. The region of interest cropped is sometimes used as a pre-processing step is used for both deep learning and machine learning approaches [17]. The pre-set unique features from these images are extracted for the machine learning algorithm. Auto-feature extraction is performed by the deep learning models. The learning is performed on labeled datasets, annotated for classification and recognition. The models developed through this approach are supervised by comparing the annotation with the predicted output.

The dataset being the key to performance and validity of the algorithms for the given task requires rigorous labeling and reviewing. The images and/or video captured are that of varied illumination, background, and views making the data challenging to learn [14]. With the region of interest extracted and labeled with key significant features extracted, there can be an improvement in learning as seen in many states of art concerning segmentation and classification. Instance segmentation is widely used in tracking where it forms part of a pipeline for segmentation. Region of interest (ROI) segmented with each instance of that specific segment can be marked and identified enabling not just detection but also tracking of individual objects in a scene[**?**]. With this perspective presented here is a robust vehicle model identification system using instance segmentation via deep learning models where the license plate and the frontal part of the car where the significant features exists are segmented for car make and model classification.

The requirement for robust vehicle identification lies in the need for public safety and security. Accuracy and real-time requirements are the prime concerns for this application. Privacy is one other element that requires to be identified in instance segmentation.

In this context, proposed here is a multi-class instance segmentation model for vehicle make and model recognition clubbed with license plate recognition presented in 1 Prominent tasks like recognition and classification are popular with this approach. However, they need a multi-step approach for vehicle frontal car segmentation and then classification. We propose a single shot segmentation network that not just identifies the vehicle make and model under varying conditions but also precedes it by segmenting the frontal part of the car as a single instance which is essential for individual unique identification and tracking. A region of interest labeled dataset for instance segmentation and a car make and license plate identification model using deep learning.

Further, we explore instance segmentation for unique vehicle identification. Instance segmentation can handle multi-detection and perform segmentation. The segmented part is then classified in a one-shot process. For the purpose of enhanced privacy and more accurate identification, an existing dataset is modified. Polygonal annotations are used that capture the curvature of the frontal part of the vehicle.

A higher accuracy for the same task on the same dataset is achieved. The inference time for the two approaches is reduced as identification of vehicle type and license plate is performed simultaneously. To improve the dataset for class imbalance data augmentation is

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 December 2022 doi:10.20944/preprints202212.0475.v1

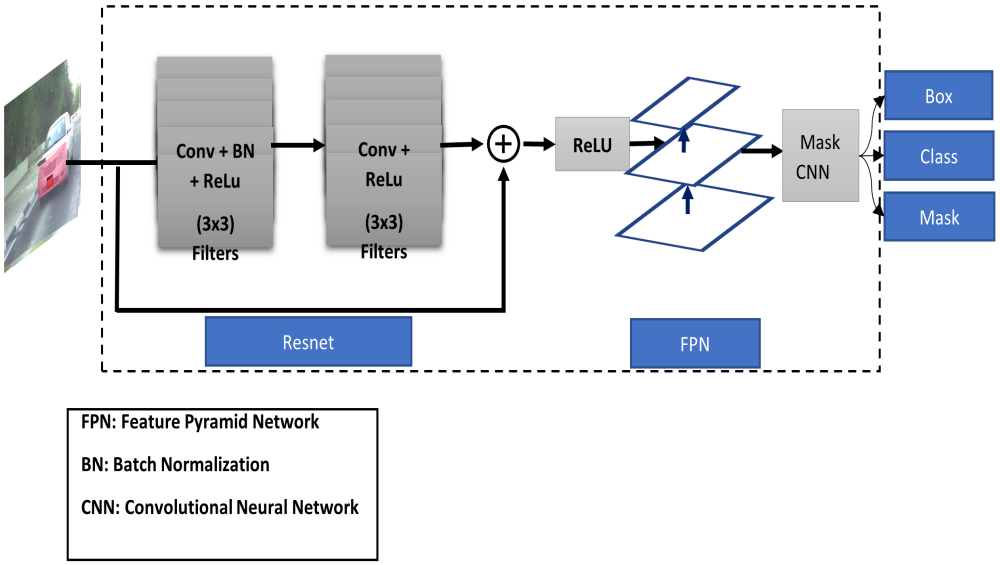

Figure 2. Mask RCNN with FPN architecture

performed in different representations and is evaluated for the same. Instance segmentation adds an ID to each unique vehicle enabling re-identification by tracking the tags. This produces a robust and accurate model for identification of vehicles in traffic, security sensitive roads and entrance of high security areas.

The contributions of this paper are as follows:

- An instance segmentation model for vehicle recognition through segmentation and classification. A single model for identifying a vehicle and identifying the make of the model with license plate.
- Achieving higher mAP of detection with a deformed convolutional network with small dataset augmented by mosaic tiling method.
- Analysis of several augmentation techniques and its effect on the recognition and detection of vehicle make identification using feature pyramid network and a deep residual network and deformed deep residual network.

2. Literature Review

Vehicle recognition is a widely researched area in the field of computer vision categorizing itself in different tasks like vehicle make and model recognition, vehicle license plate recognition and vehicle classification and vehicle re-identification. Each task is performed individually or consecutively. Application of this comes in requirements of traffic regulation systems, smart city automation, public security and even non civilian use cases. In this paper we take into perspective the requirements of a private and efficient automated vehicle make recognition system. Recent literature in this domain solves the challenges of diversity in dataset with multiple large scale datasets with large number of classes[19,24]. Further enhancing security several datasets focus on the parts and frontal area of the car enabling more fine-grained classification. In addition, datasets are varied in terms of illumination, exposure, and even environment. Large scale datasets were utilized in classifying vehicle make and model by detecting its parts specifically the frontal part which provides distinct features for vehicle classification. Of the latest in frontal image dataset was a large-scale fine-grained dataset, with diversity in scale from 103 classes. The dataset was annotated for make, model and year of manufacture providing a hierarchical representation of the vehicle. High resolution images with high quality were presented. The dataset was trained on CNN based methods. Several baseline methods have been utilized for vehicle classification including large scale models like Resnet-50. Further baseline analysis with Alexnet, VGG-16

 doi:10.20944/preprints202212.0475.v1

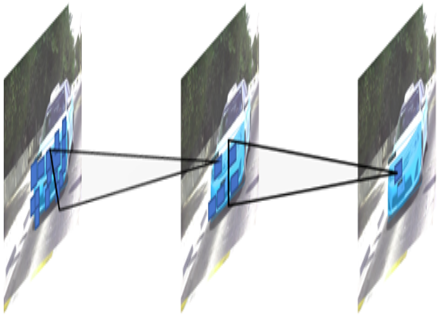

Figure 3. A deformable convolution operation performed on a vehicle instance.

and VGG-19 were performed. Each producing and accuracy above 85% thus being robust for classification[**?**].

Changing vehicle ecosystem involving new manufacturers and new models is leading an open research domain in this field. There's a requirement however for segmentation datasets annotated in polygonal format capturing enhanced contextual features of a vehicle that is not available as of now. With the aim of privacy in the perspective of application to public security, utilized in this paper is a dataset from [**?**] for instance segmentation of the frontal part of the car which includes, segmentation, detection, and classification.

Classification of make is performed using traditional rule-based approaches which are dominant in this field due to the popularity of the problem. Local and global cues were utilized for classification in several approaches. Structural and edge-based features were also a common pick. Further, machine learning was performed with these features to enhance classification. With the feature extraction techniques, edge based feature extractors like HOG and Harris corner detectors performed significantly well for detecting parts of the car like the logo, the grille and the headlights[16]. Robust feature detectors from key points like that of SIFT and SURF were employed in several state of art. Adding to these features, Corner detectors and line detectors like Hessian matrix and DoG were implemented, producing considerably higher accuracy for smaller number of classes in [**?**]. With larger number of classes, they fail to produce similar accuracy. Further, a bag of feature or Bag of Words approach was implemented with Feature detectors for unsupervised clustering producing a histogram of features for matching [18]. A typical feature detector algorithm accompanies a matching technique like, hamming distance, euclidean distance, cosine similarity for identifying similar vehicles for recognition and classification. This is further used for re-identification task.

Naïve bayes[**?**], SVM[21], LBP[21], and KNN were common machine learning algorithms used for vehicle make and model classification. CNN, used for vehicle make and model classification involve transfer learning on prominent pretrained models like that of Alexnet, VGG, Resnet, and mobilenet [15]. Adding to this modified CNN networks were introduced such as residual squeezenet [22] which produced a higher rank-5 accuracy of 99.38. Segmentation was applied as a pre-processing step to remove background. Compound scaling approach was employed on Efficient net pretrained on ImageNet for classification for the purpose of presenting an app for vehicle make and model classification. Unsupervised deep learning techniques such as auto-encoders were also utilized for this purpose [21].

With each model producing different features automatically generated through CNN based approached or engineered through edge descriptors or geometrical descriptors, there's a need for higher accuracy for a real time use cases. The cropped region of the frontal part of the car is used for identification in most cases. A segmentation of the car parts are also employed such as in [21].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 December 2022 doi:10.20944/preprints202212.0475.v1

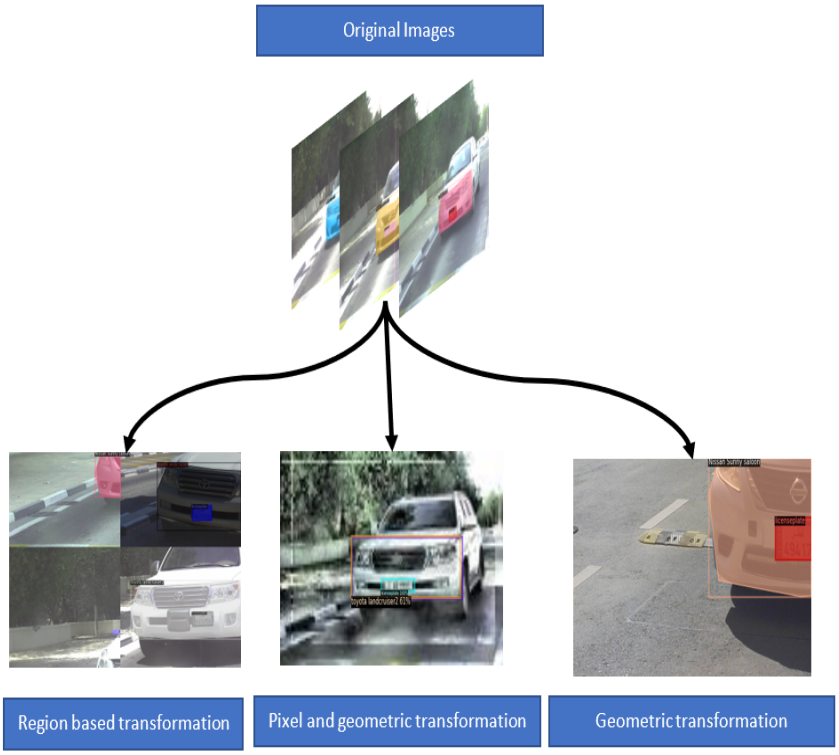

Figure 4. Augmentation techniques

Segmentation approaches are often used for removing the background and extracting the vehicle, later classifying the vehicle [23]. In a real-time use case, cropped images should be generated from an image that will later be used for part detection. Almost all approaches necessitate an extra step for vehicle detection, which adds to the time complexity. As a result, a one-step approach to vehicle identification is required. License plate detection adds to the vehicle's unique features, which are then added to the identification system for re-identification of the vehicle's unique id tagging. As a result, a robust model is required that can detect the region of interest and classify, identifying each instance of the vehicle's make.

We consider this challenge in this paper and propose instance segmentation for vehicle identification via segmentation and classification.A two-stage approach for feature extraction using FPN (a feature pyramid network produced by multi-scale feature extraction) and classification using Mask RCNN is utilized in this paper, and further experimentation is performed on a modified CNN to improve the performance of the network. Image augmentation techniques are explored for the purpose of improving an existing dataset.

3. Methods

Convolutional neural networks have been the key stone of computer vision applications. They are the most commonly used types of artificial neural networks. Convolutional operations applied to neural networks enable better feature extraction and classification. Convolutional neural networks have evolved based on the requirements of accuracy, generalization and optimization problems. In order for generalization and domain adaptation, lead to rise of several large-scale models trained on large scale data are present. Large scale data is trained on these networks which can be further adapted to other applications. Examples of convolutional neural networks being Alex net[26], Lenet [27], Resnet [29], Google-net [28], Squeeze-net [25] and so on. In this paper, we utilize Resnet which is a deep residual network consisting of multiple CNN layers. It extracts deep features and with its residual skip connections, the network is efficient in solving the vanishing gradient descent problem.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 December 2022 doi:10.20944/preprints202212.0475.v1

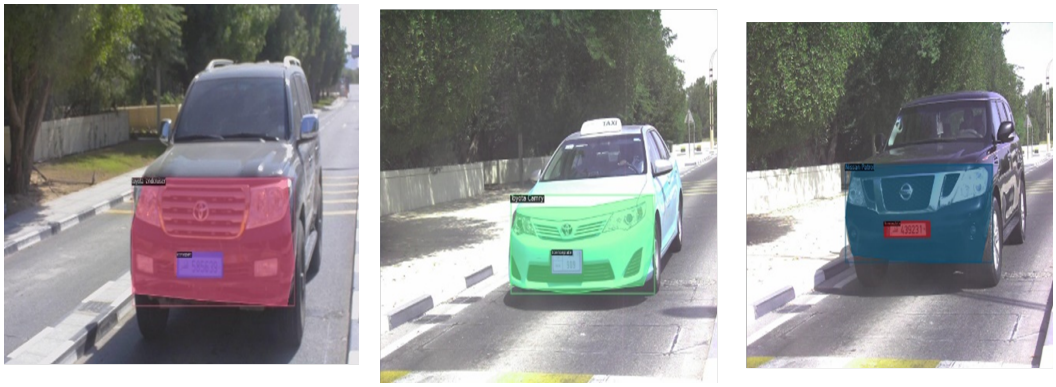

Figure 5. Dataset images .

Convolutional neural networks comprise of four key features which include weight sharing, local connection, pooling and a large number of layers [9]. The layers include the convolutional layer that perform the convolutional operation on small local patches of the input where a given input x with a filter f will produce a feature map of x. The convolution operation for the whole image is computed by the following:

Y_n = $\Sigma_{k=0}^{N-1}(x\_k)(f\_(n-k))$

where x, f, and N are the input image, filter, and the number of elements in x respectively. The output vector is represented by Y.

This is followed by activation function such as tanh, sigmoid and ReLU[30]. The activation functions introduce non-linearity into the network. The sub-sampling layer that are the pooling layers reduce the feature map resolution leading to reduce complexity and parameters. The extracted features are mapped to the labels in the fully connected layer. All the neurons are transformed into 1D format[10]. The output of convolutional and sampling layers is mapped to each of the neurons producing a fully connected layer. The fully connected layer is spatially aware extracting locational features as well as producing high level complex features. The result of this is linked to the output layer which produces output using a thresholding process. A final dense layer is sometimes used having same number of neurons as classes in case of a multi-class classification. A softmax activation function maps all the dense layer outputs to a vector producing a probability of each class.

Accuracy of this prediction is measured by its loss function where the result is compared to that of the ground truth or labelled data. A common loss function used is the categorical cross entropy loss where

$$L(xpredi, ytruei)_{=} - \Sigma_i(xpred\_i)(logytrue\_i)$$

This setup is trained through a back-propagation technique. Hyper-parameters such as learning rate, regularization and momentum parameters are set before training process and adjusted according to brute force technique. evolutionary algorithms are further used to automate hyper-parameter tuning. During the back propagation technique, the biases and weights are updated. The loss function L as in equation 2 is required to be minimum in order to produce an accurate model. For this purpose, parameters such as kernel (filters), and biases are optimized to achieve the minimum loss. He weights and biases are updated in each network and feed-forward process is iterated with the updated weights. The model converges at the least loss.

Deep residual networks are utilized as the backbone in the method used here. Deep residual networks are large networks with skip connections that carry knowledge.

In the context of this paper, Instance segmentation is performed using CNN. Instance segmentation performs detection and delineation of each object in a given image or video. Each instance of an object is tagged with an ID enabling unique detection of every object in the scene. Instance segmentation is performed in different stages which include object detection, segmentation and classification. This is enabled by CNN models as backbones

 doi:10.20944/preprints202212.0475.v1

and feature networks with classification heads. Several backbones are proposed for this approach. In this paper, we implement Mask RCNN with a Resnet backbone and Feature pyramid network. The use of this network is justified for its accuracy in object detection and segmentation where pretrained for several large datasets has superior performance over other models. However, complexity of the model causes time complexity to increase. We further measure the trade-off of the accuracy vs the time enabling evaluation of a real time use case. Figure 2 depicts the architecture of mask RCNN with FPN.

3.0.1. Deformable convolution:

With all its advantages of convolutional neural network, the geometric structures of its building modules are fixed. Augmentation is used for transforming the images as a pre-processing step in most convolutional neural networks. Thus these transformations such as rotation and orientation are fixed by modifying the training data. The structure of the filters in the kernel are also fixed rectangular window. Pooling mechanisms produce the same size of the kernels to reduce special resolution and thus the objects in the same receptive field are convoluted and presented to the activation function. Thus only identifying objects in that scale. Deformable convolution enhances geometric transformation and scaling by introducing the a 2D offset to the grid sampling locations and thereby the convolution operation offsets from its fixed receptive location to a deformed receptive field. Adding the offset thus augments the spatial sampling locations automatically. The offsets are added after the convolutional operation. The convolution operation is visualized in 3.

Further to enhance detection at lower levels, image pyramids are computed building a feature pyramid network. The object or segmentation area is scaled over different position levels in the pyramid. A proportionally sized feature maps at multiple levels are generated from a single input. Cross scale correlation is generated at each block to generate a fusion of these features. FPN's are used with CNN's as a generic solution for building feature maps. A bottom-up approach or top-down approach is used to produce a feature map. In terms of deep residual networks, the feature activation outputs are produced at each stages' last residual block. Mask R-CNN is a region-based CNN that performs object detection and classification with mask generation. The object detection is performed on a region on interest and evaluation was based on this region of interest. A multi-task loss is sampled on the Region of interest as the total of classification loss, object detection loss that is the bounding box loss and mask loss.

Complex hierarchical features are extracted from images. With extensive evaluation, the models are susceptible to overfitting. Regularization techniques are required to improve this overfitting.

3.0.2. Data Augmentation

Augmentation techniques are often applied to reduce this overfitting, that includes image transformation such as scaling, translation, rotation and random flipping. It not only increases the data size but also provides a diversity of representation. The augmentation techniques can be divided into pixel level data augmentation, region-based augmentation and geometric data augmentation. Pixel based augmentation techniques include changes in pixel values. Adding contrast, brightness or color changes the pixel intensity of the image. Regional augmentation includes that of creating masks of the required region. Motion blur and cutout are common techniques used for region-based augmentation. Geometric transformations are also applied to the data that include flipping, reflection, rotation, cropping etc. In this paper we setup the data to augment at different levels that include geometric transformation and region-based transformation. This not only enhance the dataset but also improves the datasets diversity. One particular approach used in this model is mosaic tiling method proposed in [**?**], where different training images, in this case 4, are taken in different context and stitched into one image performing a sort of mosaic tiling. Random cropping is performed on the image to reduce it to the original training image size. 4is an illustration of mosaic tiled images of the QU dataset.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 December 2022 doi:10.20944/preprints202212.0475.v1

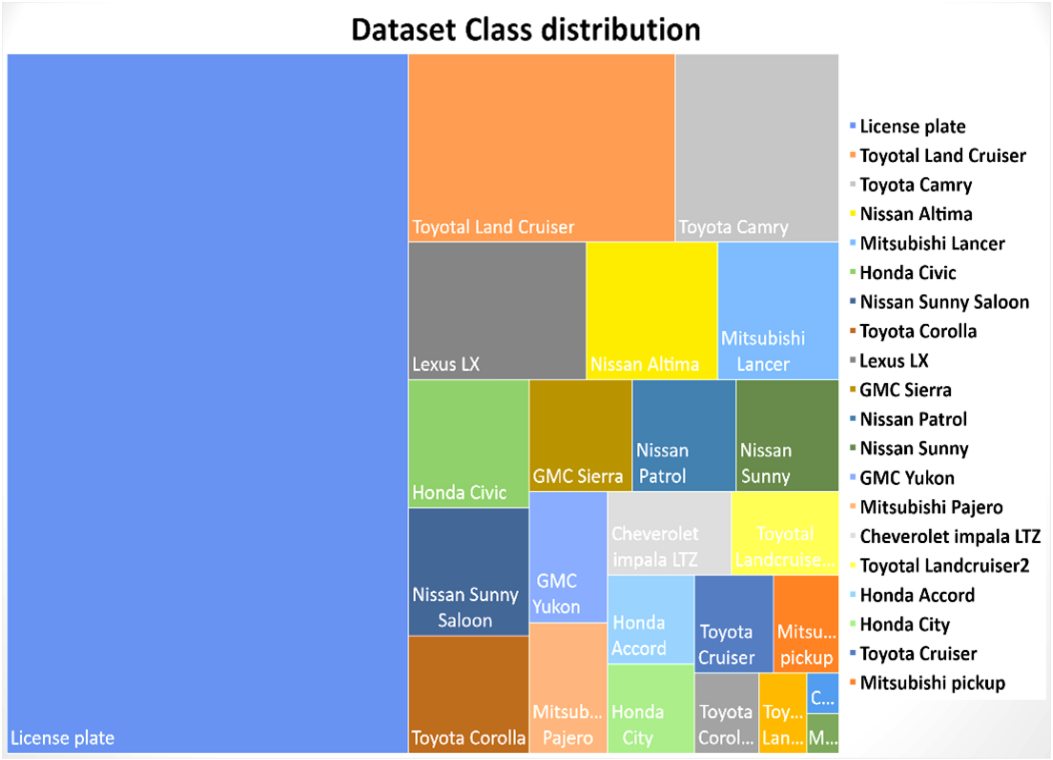

Figure 6. Dataset Distribution.

4. Experimental Setup

The setup of this network involves three layers. The vehicle with the mask is fed as training data. The data is augmented in three formats separately based on geometric augmentation and pixel-based augmentation. The transformed data is taken as the testing data and is then trained on a Mask RCNN-FPN network. Further, experiment was performed on Mask RCNN-FPN by deforming the convolutional layers. Resnet-101 and Resnet-50 are used as feature extractor backbones for performing baseline assessment on the dataset. The setup is as shown in the figure 2. The experiment was performed on Intel(R) Xeon(R) CPU @ 2.30GHz using GPU instance on an Ubuntu machine.

4.1. DATASET:

An existing dataset was modified for instance segmentation by creating polygonal bounding boxes of the frontal part of the vehicle to capture not just the frontal dashboard but also the curvature of the vehicle. The dataset contains 12 makes of vehicles taken in difference variations of camera exposure during extremely sunny weather to that of evening sunset. The dataset is bit imbalanced and so augmentation was performed to improve the data count. The vehicle distribution is as shown in Table (1). In addition, license plate is treated as a single class having a rectangular bounding box. 5 is a example of the vehicle with their annotations. A total of 225 images were split for training, testing and validation with the 157 images for training, 44 images for validation and 24 images for testing with a 70-20-10 for the original format. The classes are very imbalanced and require further augmentation. The image below displays class distribution of the dataset. This dataset contains vehicles that belong to the middle east region specifically Qatar.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 December 2022 doi:10.20944/preprints202212.0475.v1

Figure 7. Mosaic-Tiled Augmentation.

Table 1. Classification accuracy and detection acccuracy using mAP with latency

Model	Lr	Fast_rcnn/cls_accuracy	mAP	Time
Mask RCNN+RESNET-50+FPN	3x	0.992	98.772	136 ms
Mask RCNN+RESNET-101	3x	0.996	99.670	310 ms
Mask RCNN+RESNET-50	1x	0.992	99.670	316 ms
Mask RCNN+RESNET-50+FPN(DCONV)	1x	0.984375	90.747	161.81 ms

Table 2. Ablation study with different backbones and deformable convolution

Model	Model	AP	AP50	AP75
Mask RCNN-DCONV	RESNET-50 + FPN	79.648	96.337	94.350
Mask RCNN-DCONV	RESNET-50 + FPN	74.185	90.747	89.121
Mask RCNN	RESNET-50 + FPN	80.213	98.772	95.950
Mask RCNN	RESNET-101	73.621	88.219	86.265
Mask RCNN	RESNET-50	80.206	99.670	98.730

The experiments were conducted by augmenting the dataset to mimic different camera orientations and noise parameters. An evaluation of both original dataset and partly augmented dataset was performed. Augmentation parameters included in pixel and geometric based include exposure and resizing with auto-orientation, noise, and rotation. Further, patch-based augmentation which is geometric augmentation. The third type of augmentation was mosaic tiled approach. The dataset with annotation is available at [**?**]. Figure 7 is an example of data augmentation performed on the dataset and the Figure **??** shows the distribution classes across the whole dataset.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 December 2022 doi:10.20944/preprints202212.0475.v1

4.2. Performance Metrics:

To calculate the average accuracy, precision and recall must be computed for each image. TP(true positive), FP(False positive), FN(false negative) and TN(true negative) are metrics used for precision and recall. Equation 1, 2 and 3 compute the accuracy, precision and recall respectively.

$$Accuracy = \frac{Correctpred.}{Totalpred.} = \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{1}$$

$$Precision = \frac{Truepositive}{Predictedpositive} = \frac{1TP}{(TP + FP)} \tag{2}$$

$$Recall = \frac{Truepositive}{Actualpositive} = \frac{TP}{(TP + FN)} \tag{3}$$

mAP: mean Average Precision per class Average precision (AP) measures how well the model classifies each class, while mean average precision(mAP) measures how well the model classifies for all the given test dataset. It is a measure of accuracy of identification. It evaluates the performance of the model by averaging the precision under the IoU (intersection over union) with a threshold of 0.50 to 0.95. AP is calculated in each point in the threshold.

The average precision (AP) is used to evaluate the experimental performance which is calculated by averaging the precision under IoU (intersection over union) thresholds from 0.50 to 0.95 with a step of 0.05. For different queries, the evaluation metrics are APS, APM, APL, AP50, AP75, and mAP. Subscripts "S," "M," and "L" refer to "small," "medium," and "large," respectively. Subscripts "50" and "75" represent the IoU thresholds of 0.5 and 0.75, respectively. The mAP is the mean AP for each experiment.

Inference time:

The inference time is measured by the time taken to classify and generate a mask for a single input. In the context of this approach, it will be time taken to classify and generate masks for a single frame of a video.

5. Results and Discussion

Several experiments were conducted on different augmentation methods on the dataset. Resnet-50 backbone was used for the deform-able receptive field-based Mask RCNN. With a batch size of 2, the experiments ran for 1000 iterations and used a pretrained Resnet backbone on COCO dataset. Evaluation was performed using the COCO trainer module. The results without segmentation are listed in the table 1 and an ablation study based on difference backbones and feature extraction are tabulated in table 4 with the original dataset size, resolution, and clarity.

For a varied analysis, different baselines were experimented on for the purpose of evaluation and identifying the trade-off in the reliability and accuracy of an instance segmentation approach for the purpose of vehicle recognition. Mask RCNN was used as baseline with a Resnet-50 backbone with Feature pyramid network

further modelled with a Resnet-101 backbone with Feature Pyramid Network. The original dataset was augmented in multiple methods to improve the dataset description. The results of the experimentation with original dataset is displayed in Table 1. The table describes the classification accuracy of mask RCNN with that of instance segmentation accuracy with the mean average precision metric. The execution time for inference of a single image from the test set is also presented. The resent -50 back bone without FPN with base RCNN produces a high mAP of 99.670. Although Resnet-50 backbone with FPN is hypothesized to produce higher accuracy, it lags 1% but produces faster inference with 174ms faster than base-RCNN. With further experimentation on the CNN module with a deformed convolutional operation the accuracy dropped to 90% which is significantly lesser than expected. This could be due to the added complexity and generalization of the network. It can be noted that the models are inferred on a test set with imbalanced

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 December 2022 doi:10.20944/preprints202212.0475.v1

data and thus not reliable for certain classes. With class wise precision, it can be noted that the largest class, the license plate has the poorest accuracy, license plate covers a smaller area and is similar in semantic to rectangular shapes which can be a reason for the poor performance. Class wise performance is depicted in Figure 8.

Table 3. Ablation study on Data Augmentation

Augmentation Type	(Train-test-split)#	Model	Backbone	AP	AP50	AP75
Resize+expo.+rot.	471-44-24	MaskRCNN-DCONV	RESNET-50 + FPN	65.748	81.708	77.517
		MaskRCNN	RESNET-101	70.989	88.633	85.148
		MaskRCNN	RESNET-50	59.502	85.189	67.677
Full Augmentation	460-44-24	MaskRCNN-DCONV	RESNET-50 + FPN	66.780	83.101	75.029
		MaskRCNN	RESNET-101	49.585	66.776	58.586
		MaskRCNN	RESNET-50	60.163	77.906	73.954
Patch input	628-176-96	Mask RCNN-DCONV	RESNET-50 + FPN	52.475	74.535	64.246
		Mask RCNN	RESNET-101	71.569	88.176	84.842
		Mask RCNN	RESNET-50	52.186	74.393	59.095
Mosaic Based	471-44-24	Mask RCNN-DCONV	RESNET-50 + FPN	87.698	99.406	98.900
		Mask RCNN	RESNET-101	83.933	99.568	99.103
		Mask RCNN	RESNET-50	82.463	99.637	98.121

Table 4. Comparison with existing literature

Method	Model	Classification accuracy
[17]	SIFT + DoG	74.63%
Ours	MaskRCNN+ FPN + Resnet-50	99.2%

The test data is either over-represented or under-represented and thus needs to be balanced for a reliable result. Thus, multiple augmentation techniques are performed to improve data representation. Three types of augmentation approaches are utilized for this task. The following table describes the results and the approaches used. A large network and smaller network were tested to evaluate the impact of augmentation on data size and the accuracy of the model. The table below describes the results of each augmentation type on baseline models. The inference from the table is clear that mosaic augmentation performs considerably better than any other augmentation type. However, it fails to surpass images with same resolution. The patch based augmentation has very low inference than expected even though the number of images increases. This could be because of class empty patches in the dataset as each class is represented once in the original image.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 December 2022 doi:10.20944/preprints202212.0475.v1

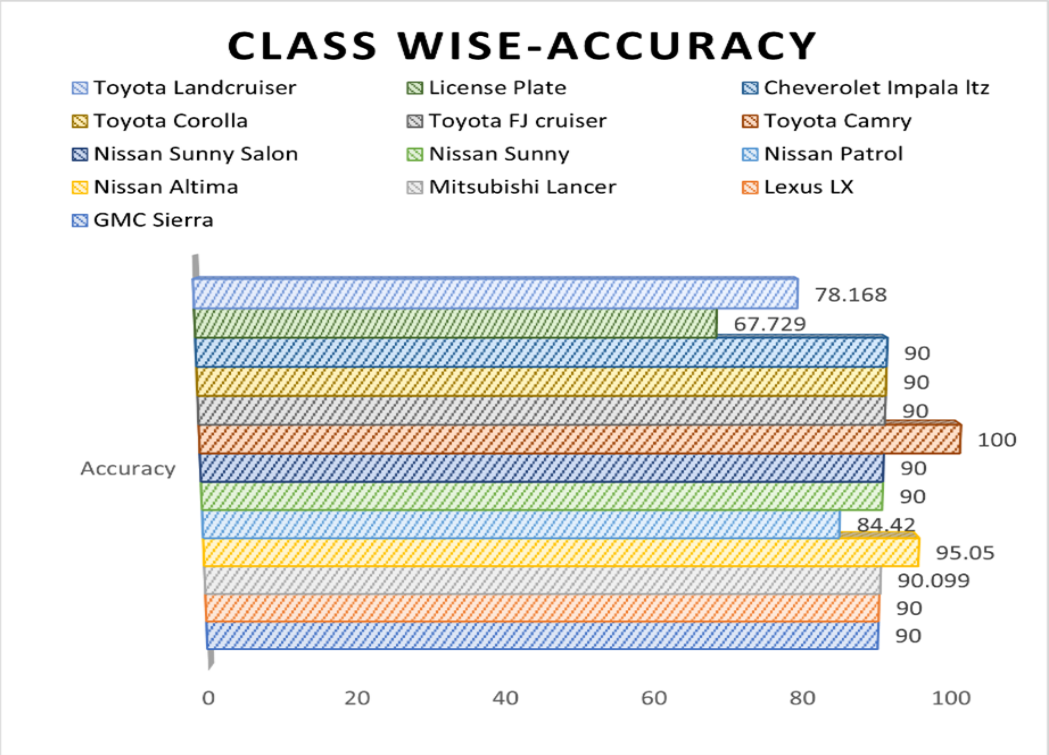

Figure 8. Class-wise accuracy based on mask RCNN-Resnet-50 results.

With per class evaluation, each class performed well on every model achieving an average of around 80%. However, License plate detection was a challenge in many models with 62.813 as the highest mAP compared to all the other networks. The number of images didn't have an impact on the performance of this class, which may be attributed to its reduced size of the license plate and its location in image with respect to models like Lexus. The figure 8 is per class result of mask RCNN with Resnet50 backbone.

5.0.1. Benchmarking:

In terms of bench-marking existing literature, the results in terms of accuracy using the existing dataset in terms of classification is given in the table **??**. The table presented shows an incredibly significant increase in accuracy compared to traditional methods using SIFT and DoG. The notable change in the model complexity and the technique produce the difference in these parameters. Distinct features are extracted globally compared to the constant local feature points on the dataset. With the same dataset a considerable increase in recognition accuracy on the test data. Although it out-stands other models, it can be seen from the figure above that classes with low number of images were not part of the test data. An imbalance is noted.

6. Conclusion

Instance segmentation of vehicular frontal region is an effective tool for vehicle classification and identification. Existing techniques requires multiple techniques to identify the vehicle, segment and then identify the make and model from this data using multiple algorithms or a separately trained network for each task. In this approach all tasks are achieved with one model. Time complexity is measured and the approach that took less execution time was mask RCNN with resnet-50 and feature pyramid network with 136ms. With an enhanced dataset with instance segmentation and further data augmentation of performance an overall evaluation technique is presented. However, new, and latest models of vehicles need to be added to the data and imbalance of dataset improved for further improvement. Further, evaluation is required for a light weight model like that of center

mask[34] which is an anchor free approach that can further improve the inference time. The instance produced from this model can be further used for re-identification as each unique instance is created for each vehicle per model. Privacy is further advanced with processing proposed in a blockchain network rather than a centralized storage as each instance of the frontal part of the vehicle can be saved rather than the whole image itself. Thus, securing the privacy and reliability of automatic vehicle recognition system is achieved.

Author Contributions: For research articles with several authors, a short paragraph specifying their individual contributions must be provided. The following statements should be used "Conceptualization, Najmath Ottakath and Somaya Al Maadeed; methodology, Najmath Ottakath.; software, Najmath Ottakath.; validation, Najmath Ottakath and Somaya Al Maadeed.; data curation, Najmath Ottakath; writing—original draft preparation, Najmath Ottakath ; writing—review and editing, Najmath Ottakath and Somaya Al Maadeed.; visualization, Najmath Ottakath; supervision,Somaya Al Maadeed; project administration, Somaya Al Maadeed; funding acquisition,Somaya Al Maadeed. All authors have read and agreed to the published version of the manuscript.", please turn to the CRediT taxonomy for the term explanation. Authorship must be limited to those who have contributed substantially to the work reported.

Funding: This research work was made possible by research grant support (QUHI-CENG-22/23-548) from Qatar University Research Fund in Qatar.

Institutional Review Board Statement: "Not Applicable"

Data Availability Statement: part of the experimental and annotated data is available at https://drive.google.com/file/d/1PnbsDMki2Abm81P-y7BoPbbcOuJUelqf/view?usp=share_link.

Conflicts of Interest: Declare conflicts of interest or state "The authors declare no conflict of interest." Authors must identify and declare any personal circumstances or interest that may be perceived as inappropriately influencing the representation or interpretation of reported research results. "The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results".

Abbreviations

The following abbreviations are used in this manuscript:

CNN	Convolutional Neural Network
DCONV	Deformed Convolution.
TL	Transfer learning.
RESNET	Residual Network.
FPN	Feature Pyramidal Network.

References

1. Elharrouss, O., Al-Maadeed, S., Subramanian, N., Ottakath, N., Almaadeed, N., Himeur, Y. (2021). Panoptic segmentation: a review. arXiv preprint arXiv:2111.10250.
2. Himeur, Y., Al-Maadeed, S., Kheddar, H., Al-Maadeed, N., Abualsaud, K., Mohamed, A., Khattab, T. (2023). Video surveillance using deep transfer learning and deep domain adaptation: Towards better generalization. Engineering Applications of Artificial Intelligence, 119, 105698.
3. Elharrouss, O., Almaadeed, N., Al-Maadeed, S. (2021). A review of video surveillance systems. Journal of Visual Communication and Image Representation, 77, 103116.
4. Akbari, Y., Almaadeed, N., Al-Maadeed, S., Elharrouss, O. (2021). Applications, databases and open computer vision research from drone videos and images: a survey. Artificial Intelligence Review, 54(5), 3887-3938.
5. McCann, J., Quinn, L., McGrath, S., Flanagan, C. (2022). Video Surveillance Architecture at the Edge (No. 9362). EasyChair.
6. Alshaikhli, M., Elharrouss, O., Al-Maadeed, S., Bouridane, A. (2021, June). Face-Fake-Net: The Deep Learning Method for Image Face Anti-Spoofing Detection: Paper ID 45. In 2021 9th European Workshop on Visual Information Processing (EUVIP) (pp. 1-6). IEEE.
7. Elharrouss, O., Almaadeed, N., Al-Maadeed, S., Bouridane, A., Beghdadi, A. (2021). A combined multiple action recognition and summarization for surveillance video sequences. Applied Intelligence, 51(2), 690-712.

8. Elharrouss, O., Almaadeed, N., Al-Maadeed, S., Bouridane, A. (2021). Gait recognition for person re-identification. The Journal of Supercomputing, 77(4), 3653-3672.
9. Li, Z., Liu, F., Yang, W., Peng, S., Zhou, J. (2021). A survey of convolutional neural networks: analysis, applications, and prospects. IEEE transactions on neural networks and learning systems.
10. Akbari, Y., Britto, A.S., Al-Maadeed, S. and Oliveira, L.S., 2019, September. Binarization of degraded document images using convolutional neural networks based on predicted two-channel images. In 2019 International Conference on Document Analysis and Recognition (ICDAR) (pp. 973-978). IEEE
11. Akbari, Y., Al-Maadeed, S. and Adam, K., 2020. Binarization of degraded document images using convolutional neural networks and wavelet-based multichannel images. IEEE Access, 8, pp.153517-153534.
12. Mohanapriya, S. (2021). Instance segmentation for autonomous vehicle. Turkish Journal of Computer and Mathematics Education (TURCOMAT), 12(9), 565-570.
13. Ottakath, N., Al-Ali, A., Al Maadeed, S. (2021). Vehicle identification using optimised ALPR.
14. Tian, B., Morris, B. T., Tang, M., Liu, Y., Yao, Y., Gou, C., ... Tang, S. (2014). Hierarchical and networked vehicle surveillance in ITS: a survey. IEEE transactions on intelligent transportation systems, 16(2), 557-580.
15. Elharrouss, O., Akbari, Y., Almaadeed, N., Al-Maadeed, S. (2022). Backbones-review: Feature extraction networks for deep learning and deep reinforcement learning approaches. arXiv preprint arXiv:2206.08016.
16. Lu, W., Zhang, H., Lan, K., Guo, J. (2009, September). Detection of vehicle manufacture logos using contextual information. In Asian Conference on Computer Vision (pp. 546-555). Springer, Berlin, Heidelberg.
17. G. Saadouli, M. I. Elburdani, R. M. Al-Qatouni, S. Kunhoth and S. Al-Maadeed, "Automatic and Secure Electronic Gate System Using Fusion of License Plate, Car Make Recognition and Face Detection," 2020 IEEE International Conference on Informatics, IoT, and Enabling Technologies (ICIoT), 2020, pp. 79-84, doi: 10.1109/ICIoT48696.2020.9089615.
18. Das, J., Shah, M., Mary, L. (2017, August). Bag of feature approach for vehicle classification in heterogeneous traffic. In 2017 IEEE International Conference on Signal Processing, Informatics, Communication and Energy Systems (SPICES) (pp. 1-5). IEEE.
19. Yang, L., Luo, P., Change Loy, C., Tang, X. (2015). A large-scale car dataset for fine-grained categorization and verification. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3973-3981).
20. Pearce, G., Pears, N. (2011, August). Automatic make and model recognition from frontal images of cars. In 2011 8th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS) (pp. 373-378). IEEE.
21. Gao, Y., Lee, H. J. (2016). Local tiled deep networks for recognition of vehicle make and model. Sensors, 16(2), 226.
22. Lee, H. J., Ullah, I., Wan, W., Gao, Y., Fang, Z. (2019). Real-time vehicle make and model recognition with the residual SqueezeNet architecture. Sensors, 19(5), 982.
23. Wu, M., Zhang, Y., Zhang, T., Zhang, W. (2020, January). Background segmentation for vehicle re-identification. In International Conference on Multimedia Modeling (pp. 88-99). Springer, Cham.
24. L. Lu, P. Wang and H. Huang, "A Large-Scale Frontal Vehicle Image Dataset for Fine-Grained Vehicle Categorization," in IEEE Transactions on Intelligent Transportation Systems, vol. 23, no. 3, pp. 1818-1828, March 2022, doi: 10.1109/TITS.2020.3027451.
25. Iandola, F. N., Han, S., Moskewicz, M. W., Ashraf, K., Dally, W. J., Keutzer, K. (2016). SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and< 0.5 MB model size. arXiv preprint arXiv:1602.07360.
26. Zahangir Alom, M., Taha, T. M., Yakopcic, C., Westberg, S., Sidike, P., Shamima Nasrin, M., ... Asari, V. K. (2018). The history began from AlexNet: a comprehensive survey on deep learning approaches. arXiv e-prints, arXiv-1803.
27. LeCun, Y. (2015). LeNet-5, convolutional neural networks. URL: http://yann. lecun. com/exdb/lenet, 20(5), 14.
28. Ballester, P., Araujo, R. M. (2016, February). On the performance of GoogLeNet and AlexNet applied to sketches. In Thirtieth AAAI conference on artificial intelligence.
29. He, K., Zhang, X., Ren, S., Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

 doi:10.20944/preprints202212.0475.v1

30. Karlik, B., Olgac, A. V. (2011). Performance analysis of various activation functions in generalized MLP architectures of neural networks. International Journal of Artificial Intelligence and Expert Systems, 1(4), 111-122.
31. H. He, Z. Shao and J. Tan, "Recognition of Car Makes and Models From a Single Traffic-Camera Image," in IEEE Transactions on Intelligent Transportation Systems, vol. 16, no. 6, pp. 3182-3192, Dec. 2015, doi: 10.1109/TITS.2015.2437998.
32. H. M. A. Bhatti, J. Li, S. Siddeeq, A. Rehman and A. Manzoor, "Multi-detection and Segmentation of Breast Lesions Based on Mask RCNN-FPN," 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2020, pp. 2698-2704, doi: 10.1109/BIBM49941.2020.9313170.
33. Lin, T. Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S. (2017). Feature pyramid networks for object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2117-2125).
34. Lee, Y., Park, J. (2020). Centermask: Real-time anchor-free instance segmentation. In Proceedings of the IEEE/CVF conference on com