sensors

MDPI

Article

Vehicle Instance Segmentation Polygonal Dataset for a Private Surveillance System

Najmath Ottakath *,† and **Somaya Al-Maadeed** †

Department of Computer Science and Engineering, Qatar University, Doha P.O. Box 2713, Qatar;
s_alali@qu.edu.qa
* Correspondence: no1912348@qu.edu.qa
† These authors contributed equally to this work.

Abstract: Vehicle identification and re-identification is an essential tool for traffic surveillance. However, with cameras at every corner of the street, there is a requirement for private surveillance. Automated surveillance can be achieved through computer vision tasks such as segmentation of the vehicle, classification of the make and model of the vehicle and license plate detection. To achieve a unique representation of every vehicle on the road with just the region of interest extracted, instance segmentation is applied. With the frontal part of the vehicle segmented for privacy, the vehicle make is identified along with the license plate. To achieve this, a dataset is annotated with a polygonal bounding box of its frontal region and license plate localization. State-of-the-art methods, maskRCNN, is utilized to identify the best performing model. Further, data augmentation using multiple techniques is evaluated for better generalization of the dataset. The results showed improved classification as well as a high mAP for the dataset when compared to previous approaches on the same dataset. A classification accuracy of 99.2% was obtained and segmentation was achieved with a high mAP of 99.67%. Data augmentation approaches were employed to balance and generalize the dataset of which the mosaic-tiled approach produced higher accuracy.

Keywords: instance segmentation; classification; vehicle make classification; mosaic-tiled augmentation

1. Introduction

Vehicle surveillance is an essential task in public security [1]. Unique features of vehicles such as the make, model, and license plate are typically utilized for traffic surveillance [2]. With traffic cameras at every intersection, the entrances of high-security buildings, parking lots, and public places, there is an opportunity to surveil and track the traffic while monitoring the road, bringing forward a smart city perspective [3]. Images and/or videos of vehicles that are captured through surveillance 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 [4–6]. Further subtasks are performed from these approaches, such as re-identification [7], tracking, and similarity matching [8–10]. Transfer learning has been widely utilized for its computing efficiency using existing pre-trained models for video surveillance [11]. 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 another element that is a requirement in public surveillance.

To achieve this objective, surveillance studies of vehicles have used machine learning and deep learning models applied to vehicle data to infer the make, model, and license plate region [12]. 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 [7] the vehicle at another location.

In the context of cars, the car's make is most prominently defined by the frontal view of the car [13]. The region of interest can be extracted from this view to identify the car's

check for updates

Citation: Ottakath, N.; Al-Maadeed, S. Vehicle Instance Segmentation Polygonal Dataset for a Private Surveillance System. *Sensors* **2023**, *23*, 3642. https://doi.org/10.3390/s23073642

Received: 31 January 2023
Revised: 19 March 2023
Accepted: 20 March 2023
Published: 31 March 2023

make and model. This enables a better representation of the uniqueness of the car. Further, the license plate can also be extracted, which can be fed to an ALPR (automatic license plate recognition) system for digit recognition, enhancing the identification of the vehicle.

In computer vision, region of interest extraction has been a task accomplished by segmentation. The cropped region of interest is sometimes used as a pre-processing step for both deep learning and machine learning approaches [14]. A pre-set set of unique features from these images is extracted for the machine learning algorithm, whereas auto-feature extraction is performed by deep learning models.

The data presented for learning, being key to the performance and validity of the algorithms for a given task, requires rigorous labelling and reviewing. The images and/or videos captured are those of varied illumination, background, and views, making the data challenging to learn [15]. With the region of interest extracted and labelled with key significant features, there can be an improvement in learning, as seen in many state-of-the-art methods concerning segmentation and classification [16].

Instance segmentation is a task used in tracking. A 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 [17]. In this context, utilized here is a multi-class instance segmentation for vehicle make and model recognition clubbed with license plate recognition, as presented in Figures 1 and 2. Typical make and model identification techniques need a multi-step approach for vehicle frontal-part segmentation and then classification of the detected vehicle. In this paper, a segmentation network is proposed that not only identifies the vehicle make and model under varying conditions but also precedes it by segmenting the significant frontal part of the car as a single instance, which safeguards privacy and is essential for individual unique identification and tracking. This paper presents an unique region-of-interest-labelled dataset for instance segmentation with polygonal annotations and vehicle make classification and license plate localization using deep learning.

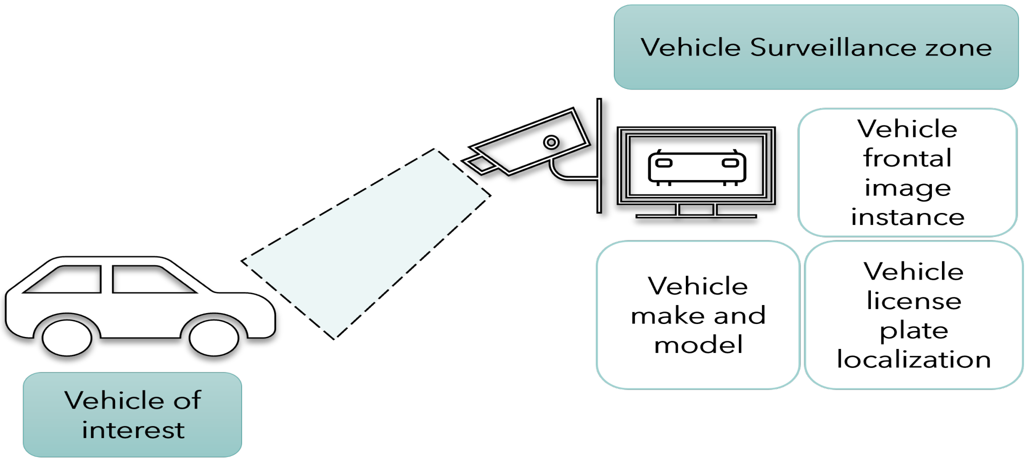
Vehicle Surveillance zone
Vehicle frontal image instance
Vehicle make and model
Vehicle license plate localization
Vehicle of interest

Figure 1. The process of vehicle surveillance at the camera for make and model classification and license plate localization.

Within this framework, 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 the vehicle type and license plate is performed simultaneously. To improve the dataset for class imbalance, data augmentation is performed in different representations and is evaluated on the same dataset. This produces a robust and accurate model for identification of vehicles in traffic, security-sensitive roads and entrances to 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 a higher mAP of detection with a deformed convolutional network with a small dataset augmented by the mosaic-tiling method.
- Analysis of several augmentation techniques and their effect on the recognition and detection of vehicle make identification using feature pyramid, deep residual and deformed deep residual networks.

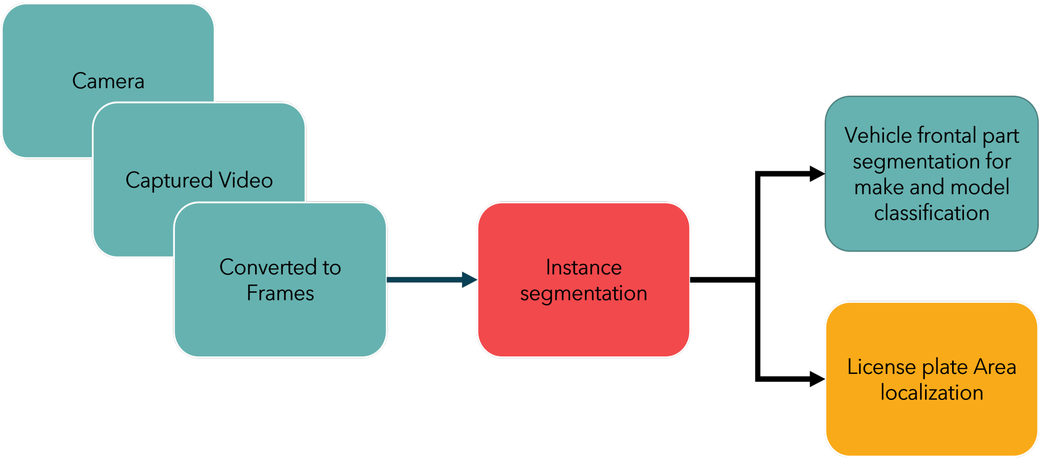
Camera
Captured Video
Converted to Frames
Instance segmentation
Vehicle frontal part segmentation for make and model classification
License plate Area localization

Figure 2. Vehicle instance segmentation technique for license plate localization along with make and model classification.

In comparison to existing literature using the same dataset in [14], this method produced higher classification accuracy, with a 25.5% increase. Further, the inference time is reduced to milliseconds. Polygonal annotation of the frontal region of the vehicle is a novel approach leading to a high mAP of 99.67% for segmentation. Thus, when compared to the full vehicle instance segmentation using the KITTI dataset, the model achieved only 92%, as demonstrated in [18].

With vehicle make classification and license plate localization achieved through instance segmentation, the goal was to discover the ability of the deep learning model to perform on a polygonal dataset. In this regard, an ablation study was conducted on a deep learning framework by modifying its backbone to measure the accuracy and time complexity of each, through which the reliability of each approach was measured. Additionally, to evaluate the dataset for generalization and reducing the imbalance, an ablation study using various image augmentation techniques was performed. With the vehicles' frontal part segmented, private surveillance was achieved.

2. Literature Review

Vehicle recognition is a widely researched area in the field of computer vision, categorizing itself in different tasks such as vehicle make and model recognition (VMMR), vehicle license plate recognition and vehicle re-identification [19,20]. Each task is performed individually or consecutively. The application of this comes with requirements of traffic regulation systems, smart city automation, public security and even non-civilian use cases [21]. In this paper, we take into consideration the requirements of a private and efficient automated vehicle make recognition system through instance segmentation.

Recent literature in this domain solves the challenges of private surveillance with dataset diversity with multiple large scale datasets containing a large number of classes [13,22]. This enhances not just privacy but also efficient vehicle recognition, with several datasets focusing on the frontal area of the car enabling more fine-grained classification. However, with similar vehicle features and diverse environments there still exists unique challenges in vehicle recognition. Changing vehicle ecosystem involving new manufacturers and new models has led to an open research domain in this field. There is a requirement, however, for segmentation datasets annotated in polygonal format capturing enhanced contextual features of a vehicle which is currently non-existent. With the aim of privacy and public security in its application, this paper utilises a dataset from [14] for instance segmentation of the frontal part of the car which includes, segmentation, detection, and classification.

Classification of vehicle make is performed using traditional rule-based approaches which are prominent in this field due to the popularity of the problem. Local and global cues have been utilized for classification in several approaches. Structural and edge-based features have also been a common pick. Further, machine learning has been performed with these features to enhance classification. With the feature extraction techniques, edge-based feature extractors, such as HOG and Harris corner detectors, have performed significantly well for detecting parts of the car such as the logo, the grille or the headlights [23]. Robust feature detectors from key points, such as SIFT and SURF, have been employed in several state-of-the-art models. In addition to these features, corner and line detectors, such as Hessian matrix and DoG (difference of gaussian), have been implemented, producing considerably a higher accuracy for a smaller number of classes [14,24]. With a larger number of classes, these models failed to produce a similar accuracy. Further, a bag-of-features or a bag-of-words approach has been implemented with feature detectors for unsupervised clustering producing a histogram of features for matching [25]. A typical feature detector algorithm accompanies a matching technique, such as hamming distance, euclidean distance, or cosine similarity, to identify similar vehicles for recognition and classification. This is further used for re-identification tasks.

Naïve Bayes [26], SVM [27], LBP [27], and KNN are common machine learning algorithms that have been used for vehicle make and model classification. CNN architecture used for vehicle make and model classification involves transfer learning on prominent pre-trained models, such as Alexnet, VGG, Resnet, and mobilenet [28]. In addition to this, modified CNN networks were introduced, such as residual squeezenet [2], which produces a higher rank-5 accuracy of 99.38%. Segmentation has been applied as a pre-processing step to remove background noise. The compound scaling approach has been employed on EfficientNet pre-trained on ImageNet for vehicle make and model classification. Unsupervised deep learning techniques such as auto-encoders have also been utilized for this purpose [27]. Apart from frontal images, recently a part-level feature extraction method where feature grouping was utilized by Lei et al. in [29] was employed to classify and recognize vehicles. This method produced an recognition accuracy of 97.7%. A genetic algorithm for feature optimization of CNN-generated features was utilized in [30]. Classification was performed using an SVM classifier. A hybrid CNN–SVM method was performed which produced an accuracy of 99.71%; however, this method failed to present license plate localization or region of interest segmentation.

Segmenting the region of interest achieves better recognition and private surveillance. In vehicle identification, segmentation approaches are often used to remove the background and extract the vehicle to classifying it [27,31]. 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 for 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, 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) [32] and classification using maskRCNN [33] is utilized in this paper. 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 the existing dataset.

3. Methods

Convolutional neural networks have been key in computer vision applications. They are the most commonly used type of artificial neural networks. Convolutional operations applied to neural networks enable better feature extraction and classification [34]. Convolutional neural networks have evolved based on the requirements of accuracy, generalization and optimization problems. The need for generalization and domain adaptation has led to a rise in several large-scale models trained on large-scale data. Large-scale data is trained on these networks which can be further adapted to other applications. Examples of convolutional neural networks include, Alex net [35], Lenet [36], Resnet [37], Google-net [38], Squeeze-net [39], and so on. In this paper, we utilize Resnet, a deep residual network consisting of multiple CNN layers. Resnet extracts deep features and with its residual skip connections, the network is efficient in solving the vanishing gradient problem [37].

Convolutional neural networks comprise of four key features which include weight sharing, local connection, pooling and a large number of layers [40]. The layers include the convolutional layer that performs the convolutional operation on small local patches of the input where a given input x with a filter f produces a feature map of x. The convolution operation for the whole image is computed by the following, as shown in Equation (1).

$$Y\_n = \Sigma_{k=0}^{N-1}(x\_k)(f(n-k)) \tag{1}$$

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 functions such as tanh, sigmoid and ReLU [41]. 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 reduced complexity and parameters. The extracted features are mapped to the labels in the fully connected layer. All the neurons are transformed into a 1D format [42]. The outputs of the convolutional and sampling layers are 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 the output using a thresholding process. A final dense layer is sometimes used with the same number of neurons as classes in the case of multi-class classification. A softmax activation function maps all the dense layer outputs to a vector producing a probability for each class.

The accuracy of this prediction is measured by its loss function where the result is compared to the ground truth or labelled data. A commonly used loss function is the categorical cross-entropy loss computed as L, as shown in Equation (2).

$$L = -\Sigma_{i=1}^{N}(y_i \ \cdot log(y\,\hat{i})) \tag{2}$$

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

Deep residual networks are utilized as the backbone. Deep residual networks are large networks with skip connections that carry knowledge. The methodology utilized in this framework performs instance segmentation using a CNN. Instance segmentation enables 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.

3.1. Deformable Convolution

With all the advantages of the convolutional neural network, the geometric structures of its building modules are fixed. Augmentation is used for transforming 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 in a rectangular window. Pooling mechanisms produce the same size kernels to reduce special resolution and thus the objects in the same receptive field are convoluted and presented to the activation function. Therefore, only objects in that scale are identified. Deformable convolution enhances geometric transformation and scaling by introducing 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 automatically augments the spatial sampling locations. The offsets are added after the convolutional operation.

Further, to enhance detection at lower levels image pyramids are computed to build a feature pyramid network. The object or segmentation area is scaled over different position levels in the pyramid. Proportionally sized feature maps at multiple levels are generated from a single input. Then, cross-scale correlation is generated at each block to generate a fusion of these features. FPNs are used with CNNs as a generic solution to build feature maps. A bottom-up or top-down approach is then used to produce a feature map. In terms of deep residual networks, the feature activation outputs are produced at each stages' last residual block.

In this paper, we implement a maskRCNN with a Resnet backbone and a FPN. The use of this network is justified due to its accuracy in object detection and segmentation when it is pre-trained on several large datasets which have superior performance over other models. However, the complexity of the model causes the time complexity to increase. Therefore, we further measure the trade-off accuracy vs. time enabling the evaluation of a real-time use case. Figure 3 depicts the architecture of the maskRCNN with FPN used for instance segmentation. The maskRCNN is a region-based CNN that performs object detection and classification with mask generation. The object detection is performed on a region of interest and then evaluated. A multi-task loss is sampled from the region of interest as a total classification loss, and object detection loss is the bounding box loss and mask loss. Complex hierarchical features are extracted from images. With extensive evaluation, the models are susceptible to overfitting; therefore, regularization techniques are required to improve this overfitting.

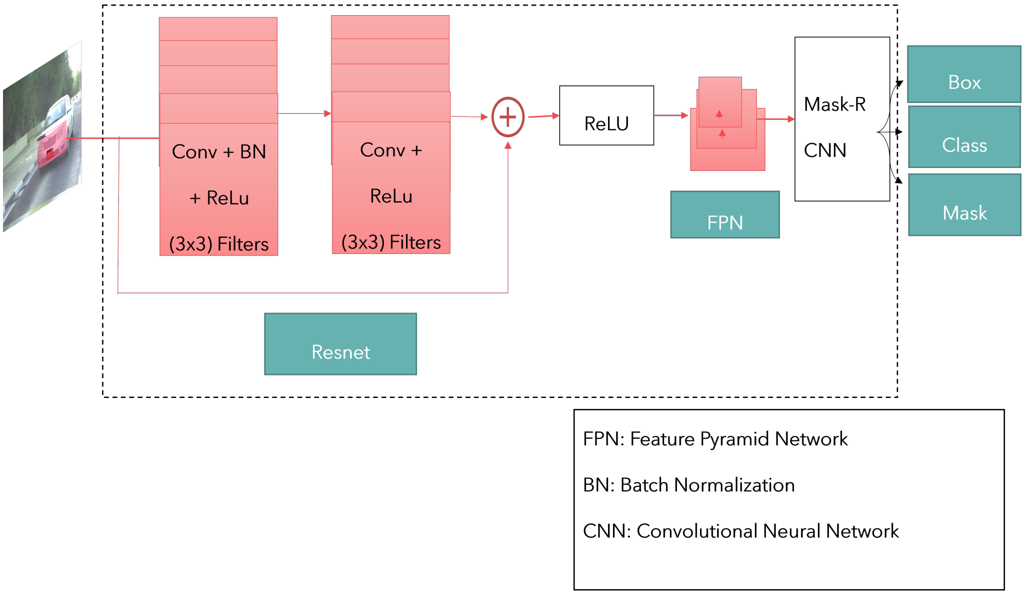

Figure 3. The maskRCNN with FPN architecture.

3.2. Data Augmentation

Augmentation techniques are often applied to reduce overfitting, this includes image transformation, such as scaling, translation, rotation and random flipping. This not only increases the data size but also provides a diversity of representation. The augmentation techniques can be divided into pixel-level data, region-based and geometric data augmentation. Pixel-based augmentation techniques include changes in pixel values. Adding contrast, brightness and colour changes the pixel intensity of the image. Regional augmentation includes creating masks of the required region. Motion blur and cut-out are common techniques used for region-based augmentation. Geometric transformations are also applied to data including flipping, reflection, rotation, cropping, etc. In this paper, we set up the data to augment at different levels that include geometric transformation and region-based transformation, as seen in Figure 4. This not only enhances the dataset but also improves dataset diversity. One particular approach used in this model is the mosaic-tiling method proposed in [43], where different training images, in this case four, are taken in different context and stitched into one image to create a sort of mosaic tiling. Random cropping is performed on the image to reduce it to the original training image size. Figure 5 is an illustration of the mosaic-tiled images of the dataset. Thus, a baseline method is used for instance segmentation and then modified and evaluated in terms of data augmentation, different feature extractors, and deformed convolution to identify the effect of each and chose the optimum configuration for vehicle instance segmentation.

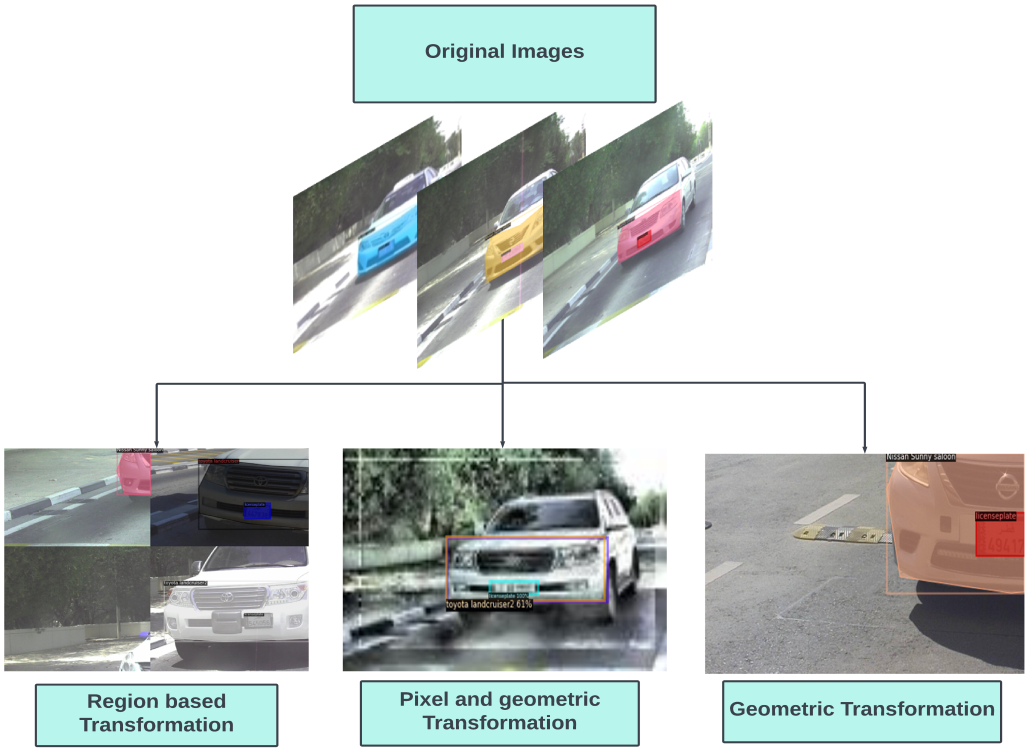

Figure 4. Augmentation techniques.

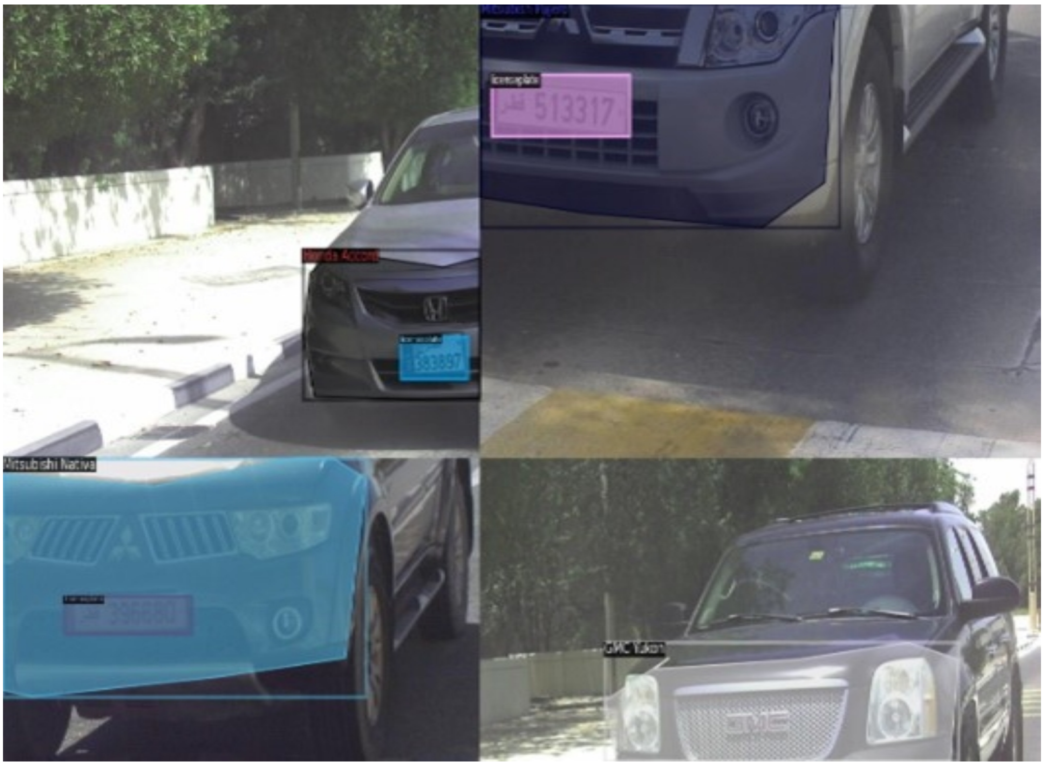

Figure 5. Mosaic-tiled augmentation.

4. Experimental Setup

The setup of this network involves three layers. The vehicle with a mask is fed in as the training data. The data is separately augmented in three formats based on geometric and pixel-based augmentation. The transformed data is taken as the testing data and then trained on a maskRCNN-FPN network. Further, an experiment was performed on the maskRCNN-FPN by deforming the convolutional layers. Resnet-101 and 50 were used as feature extractor backbones to perform baseline assessments on the dataset. All experiments were ran with a learning rate of 0.001 and 5000 iterations. The setup is as shown in Figure 3. The experiment was performed on an Intel(R) Xeon(R) CPU @ 2.30 GHz using GPU virtual instance on an Ubuntu machine (Asus, China).

4.1. Dataset

An existing dataset [14] was modified for instance segmentation by creating polygonal bounding boxes of the frontal part of the vehicle to capture the frontal dashboard and the curvature of the vehicle. The dataset contains 12 vehicles makes taken from different camera exposures during extremely sunny weather to evening sunset. The dataset is imbalanced and therefore augmentation was performed to improve the data count. In addition, the license plate was treated as a single class having a rectangular bounding box. Figure 6 shows the vehicle samples with their annotations. A total of 225 images were split for training, testing and validation with 157 images for training, 44 images for validation, and 24 images for testing (a 70:20:10 ratio from the original format). This split was utilized to match the split of the reference paper [14]. The classes were very imbalanced and required further augmentation, performed as per the methodology stated earlier. The image below displays the class distribution of the dataset. This dataset contains vehicles that belong to the middle-eastern region, specifically Qatar.

The experiments were conducted by augmenting the dataset to mimic different camera orientations and noise parameters. An evaluation of both the original dataset and partly augmented dataset was performed. The augmentation parameters included in the pixel- and geometric-based augmentation include exposure and resizing with auto-orientation, noise, and rotation. Further, patch-based augmentation was performed which is a type of geometric augmentation. The third type of augmentation was the mosaic-tiled approach. The dataset with annotation is available at [44]. Figure 5 is an example of data augmentation performed on the dataset and Figure 7 shows the distribution of classes across the whole dataset.

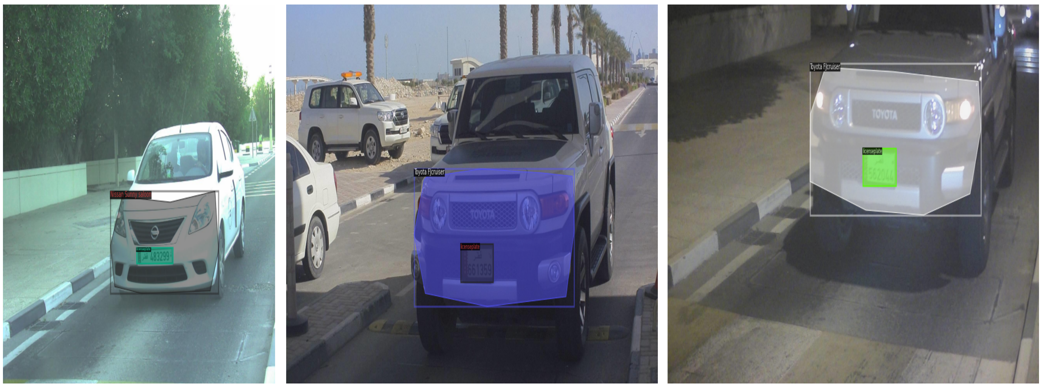

Figure 6. Dataset images.

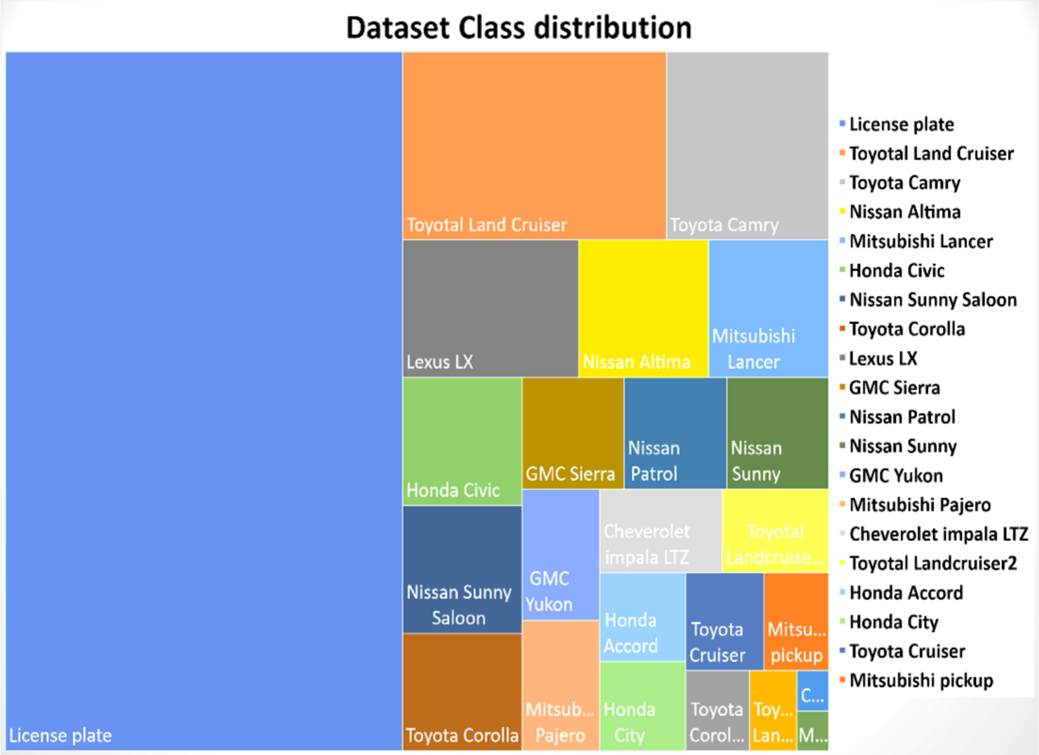

Figure 7. Dataset distribution.

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 the metrics used for precision and recall. Equations (3)–(5) were used to compute the accuracy, precision and recall, respectively.

$$Accuracy = \frac{Correct pred.}{Total pred.} = \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{3}$$

$$Precision = \frac{True positive}{Predicted positive} = \frac{1TP}{(TP + FP)} \tag{4}$$

$$Recall = \frac{Truepositive}{Actualpositive} = \frac{TP}{(TP+FN)} \tag{5}$$

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 the whole given test dataset. It is a measure of identification accuracy. It evaluates the performance of the model by averaging the precision under the IoU (intersection over union) with a threshold from 0.50 to 0.95, with a step of 0.05. The AP is calculated for each point within the threshold. 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, the inference time will be taken to classify and generate masks for a single frame of a video.

5. Results and Discussion

Several experiments were conducted for different augmentation methods on the dataset. The Resnet-50 backbone was used for the deformable receptive field-based maskR-CNN. With a batch size of two, the experiments ran for 1000 iterations and used a pre-trained Resnet backbone on the COCO dataset. The evaluation was performed using the COCO trainer module. The results without segmentation are listed in Table 1 and the ablation study based on different backbones and feature extractions is tabulated in Table 2 with the original dataset size, resolution, and clarity.

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

Model	Lr	Fast_rcnn/cls_accuracy	mAP	Time
MaskRCNN + RESNET-50 + FPN	3×	0.992	98.772	136 ms
MaskRCNN + RESNET-101	3×	0.996	88.219	310 ms
MaskRCNN + RESNET-50	1×	0.992	99.670	316 ms
MaskRCNN + RESNET-50 + FPN (DCONV)	1×	0.984375	90.747	161.81 ms

Table 2. Ablation study with different backbones and deformable convolution.

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

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. The maskRCNN was used as a baseline with a Resnet-50 backbone and FPN. Further, the maskRCNN was modelled with a Resnet-101 backbone with FPN. The original dataset was augmented with multiple methods to improve the dataset description. The results of experiments with the original dataset is displayed in Tables 1 and 2. The Table 1 describes the classification accuracy of the maskRCNN with the instance segmentation accuracy and mean average precision metric. The execution time for inference of a single image from the test set is also presented. The Resnet-50 backbone with a base RCNN without the FPN produces a high mAP of 99.670%.

Although the Resnet-50 backbone with the FPN is hypothesized to produce a higher accuracy, it lags by 1% but produces a faster inference, 174 ms faster than the base-RCNN. Further experiments on the CNN module with a deformed convolutional operation the accuracy dropped to 90%, significantly less than expected. This could be due to the added complexity and generalization of the network. It should be noted that the models are inferred on a test set with imbalanced data and thus are not reliable for certain classes. With class-wise precision, it should be noted that the largest class, license plate detection has the poorest accuracy. This poor performance may be because the license plate covers a small area and is similar to other rectangular shapes. Class-wise performance is depicted in Figure 8.

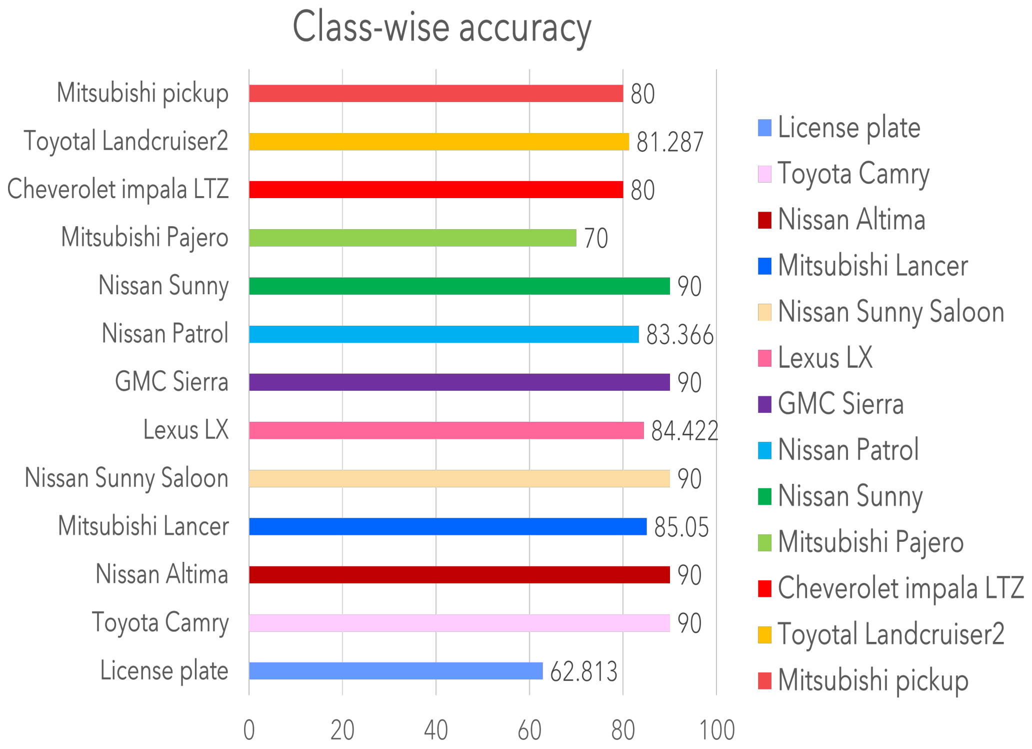

Figure 8. Class-wise accuracy based on the maskRCNN–Resnet-50 results.

The test data is either over- or under-represented and thus has to be balanced for reliable results. Thus, multiple augmentation techniques were performed to improve data representation. Three types of augmentation approaches were utilized for this task. Table 3 describes the results and the approaches used. Large and small networks were tested to evaluate the impact of augmentation on data size and model accuracy. The table describes the results of each augmentation type on the 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 the same resolution. The patch-based augmentation has a much lower inference than expected even though the number of images increase. This could be because of class-empty patches in the dataset as each class is represented once in the original image. With per class evaluation, each class performed well in every model achieving an average of around 80%. However, license plate detection was a challenge for many of the models with 62.813% as the highest mAP compared to all the other networks. The number of images did not have an impact on the performance of this class, which may be attributed to the reduced size of the license plate and its location in images with respect to models such as Lexus. Figure 8 shows the per class result of the maskRCNN with the Resnet-50 backbone with the highest accuracy for the Toyota corolla compared with other classes. Figure 9 shows a resultant image of segmentation where the frontal area of the vehicle is segmented and the make identified with license plate localization.

Table 3. Ablation study on data augmentation.

Aug. Type	(Train-Test-Split)	Model	Backbone	AP	AP50	AP75
Resize+expo.	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 Aug.	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	MaskRCNN-DCONV	Resnet-50 + FPN	52.475	74.535	64.246
		MaskRCNN	Resnet-101	71.569	88.176	84.842
		MaskRCNN	Resnet-50	52.186	74.393	59.095
Mosaic Based	471-44-24	MaskRCNN-DCONV	Resnet-50 + FPN	87.698	99.406	98.900
		MaskRCNN	Resnet-101	83.933	99.568	99.103
		MaskRCNN	Resnet-50	82.463	99.637	98.121

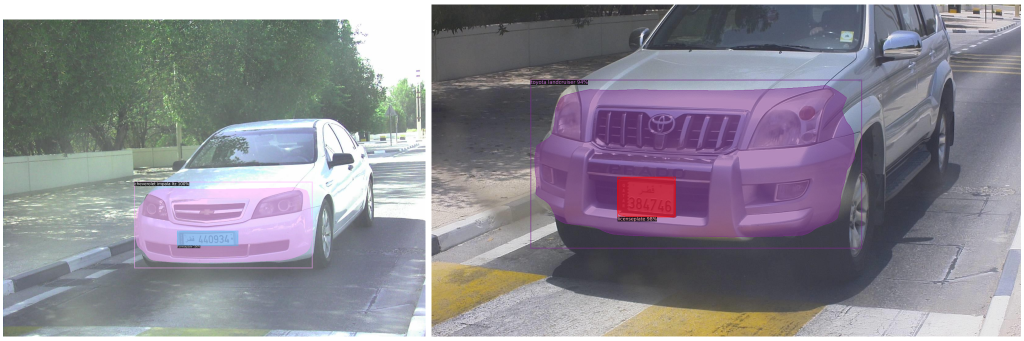

Figure 9. Resultant images of segmentation and classification using the maskRCNN with the Resnet-50 backbone.

Benchmarking

Bench-marking existing literature, the classification accuracy using the existing dataset is given in Table 4. The table shows an significant increase in accuracy compared to traditional methods using SIFT and DoG. The notable change in the model complexity and technique produce the difference in these parameters. Distinct features are globally extracted compared to the constant local feature points in the dataset. Comparing existing results on the same dataset, a considerable increase in recognition accuracy was achieved on the test data. Although it stands out from other models, it can be seen from Figure 8 that classes with a low number of images were not part of the test data. Therefore, an imbalance is noted.

Table 4. Comparison with the existing literature.

Reference	Model	Classification Accuracy
[14]	SIFT + DoG	74.63%
Ours	MaskRCNN+ FPN + Resnet-50	99.2%

6. Conclusions

Instance segmentation of a vehicle's frontal region is an effective tool for vehicle classification and identification. Existing techniques require multiple steps to identify a vehicle, segment and then identify the make and model from this data using multiple algorithms or separately trained networks for each task. In this approach all tasks were achieved with one model. Time complexity was measured and the approach that exhibited the lowest execution time (136 ms) was the maskRCNN with the Resnet-50 and FPN. With an enhanced dataset with instance segmentation and further data augmentation of the performance an overall evaluation is presented. However, new models of vehicles need to be added to the data to balance the dataset for further improvement. Further, evaluation is required for light weight models, such as the centre mask [45] model which is an anchor-free approach that can further improve the inference time. The instance produced from this model could 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 the automatic vehicle recognition system is achieved.

Author Contributions: Conceptualization, N.O. and S.A.-M.; methodology, N.O.; software, N.O.; validation, N.O. and S.A.-M.; data curation, N.O.; writing—original draft preparation, N.O.; writing—review and editing, N.O. and S.A.-M.; visualization, N.O.; supervision, S.A.-M.; project administration, S.A.-M.; funding acquisition, S.A.-M. All authors have read and agreed to the published version of the manuscript.

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

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Part of the experimental and annotated data are available at https://drive.google.com/drive/folders/1zqR1s9YiTxAfjfF213WbiH3Xc-SHPIPs?usp=sharing (accessed on 1 December 2022).

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

CNN	Convolutional neural network
D-CONV	Deformed convolution
TL	Transfer learning
RESNET	Residual network
FPN	Feature pyramidal network

References

1. Räty, T.D. Survey on contemporary remote surveillance systems for public safety. *IEEE Trans. Syst. Man Cybern. Part C (Appl. Rev.)* **2010**, *40*, 493–515. [CrossRef]
2. Lee, H.J.; Ullah, I.; Wan, W.; Gao, Y.; Fang, Z. Real-time vehicle make and model recognition with the residual SqueezeNet architecture. *Sensors* **2019**, *19*, 982. [CrossRef] [PubMed]
3. Zhang, Y.; Sun, Y.; Wang, Z.; Jiang, Y. YOLOv7-RAR for Urban Vehicle Detection. *Sensors* **2023**, *23*, 1801. [CrossRef]
4. Elharrouss, O.; Al-Maadeed, S.; Subramanian, N.; Ottakath, N.; Almaadeed, N.; Himeur, Y. Panoptic segmentation: A review. *arXiv* **2021**, arXiv:2111.10250.
5. Elharrouss, O.; Almaadeed, N.; Al-Maadeed, S. A review of video surveillance systems. *J. Vis. Commun. Image Represent.* **2021**, *77*, 103116. [CrossRef]
6. McCann, J.; Quinn, L.; McGrath, S.; Flanagan, C. *Video Surveillance Architecture at the Edge (No. 9362)*; EasyChair: Manchester, UK, 2022.
7. Elharrouss, O.; Almaadeed, N.; Al-Maadeed, S.; Bouridane, A. Gait recognition for person re-identification. *J. Supercomput.* **2021**, *77*, 3653–3672. [CrossRef]
8. Akbari, Y.; Almaadeed, N.; Al-Maadeed, S.; Elharrouss, O. Applications, databases and open computer vision research from drone videos and images: A survey. *Artif. Intell. Rev.* **2021**, *54*, 3887–3938. [CrossRef]
9. Alshaikhli, M.; Elharrouss, O.; Al-Maadeed, S.; Bouridane, A. Face-Fake-Net: The Deep Learning Method for Image Face Anti-Spoofing Detection: Paper ID 45. In Proceedings of the 2021 9th European Workshop on Visual Information Processing (EUVIP), Paris, France, 23–25 June 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 1–6.
10. Elharrouss, O.; Almaadeed, N.; Al-Maadeed, S.; Bouridane, A.; Beghdadi, A. A combined multiple action recognition and summarization for surveillance video sequences. *Appl. Intell.* **2021**, *51*, 690–712. [CrossRef]

11. Himeur, Y.; Al-Maadeed, S.; Kheddar, H.; Al-Maadeed, N.; Abualsaud, K.; Mohamed, A.; Khattab, T. Video surveillance using deep transfer learning and deep domain adaptation: Towards better generalization. *Eng. Appl. Artif. Intell.* **2023**, *119*, 105698. [CrossRef]
12. Ottakath, N.; Al-Ali, A.; Al Maadeed, S. Vehicle Identification Using Optimised ALPR. 2021. Available online: http://hdl.handle.net/10576/24527 (accessed on 20 January 2023).
13. Lu, L.; Wang, P.; Huang, H. A Large-Scale Frontal Vehicle Image Dataset for Fine-Grained Vehicle Categorization. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 1818–1828. [CrossRef]
14. Saadouli, G.; Elburdani, M.I.; Al-Qatouni, R.M.; Kunhoth, S.; Al-Maadeed, S. Automatic and Secure Electronic Gate System Using Fusion of License Plate, Car Make Recognition and Face Detection. In Proceedings of the 2020 IEEE International Conference on Informatics, IoT, and Enabling Technologies (ICIoT), Doha, Qatar, 2–5 February 2020; pp. 79–84. [CrossRef]
15. Tian, B.; Morris, B.T.; Tang, M.; Liu, Y.; Yao, Y.; Gou, C.; Shen, D.; Tang, S. Hierarchical and networked vehicle surveillance in ITS: A survey. *IEEE Trans. Intell. Transp. Syst.* **2014**, *16*, 557–580. [CrossRef]
16. Ali, M.; Tahir, M.A.; Durrani, M.N. Vehicle images dataset for make and model recognition. *Data Brief* **2022**, *42*, 108107. [CrossRef] [PubMed]
17. Mohanapriya, S. Instance segmentation for autonomous vehicle. *Turk. J. Comput. Math. Educ.* **2021**, *12*, 565–570.
18. Ojha, A.; Sahu, S.P.; Dewangan, D.K. Vehicle detection through instance segmentation using mask R-CNN for intelligent vehicle system. In Proceedings of the 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 6–8 May 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 954–959.
19. Khan, S.W.; Hafeez, Q.; Khalid, M.I.; Alroobaea, R.; Hussain, S.; Iqbal, J.; Almotiri, J.; Ullah, S.S. Anomaly detection in traffic surveillance videos using deep learning. *Sensors* **2022**, *22*, 6563. [CrossRef] [PubMed]
20. Shidik, G.F.; Noersasongko, E.; Nugraha, A.; Andono, P.N.; Jumanto, J.; Kusuma, E.J. A systematic review of intelligence video surveillance: Trends, techniques, frameworks, and datasets. *IEEE Access* **2019**, *7*, 170457–170473. [CrossRef]
21. Olatunji, I.E.; Cheng, C.H. Video analytics for visual surveillance and applications: An overview and survey. In *Machine Learning Paradigms: Applications of Learning and Analytics in Intelligent Systems*; Springer: Cham, Switzerland, 2019; pp. 475–515.
22. Yang, L.; Luo, P.; Loy, C.C.; Tang, X. A large-scale car dataset for fine-grained categorization and verification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3973–3981.
23. Lu, W.; Zhang, H.; Lan, K.; Guo, J. Detection of vehicle manufacture logos using contextual information. In Proceedings of the Asian Conference on Computer Vision, Xi'an, China, 23–27 September 2009; Springer: Berlin/Heidelberg, Germany, 2009; pp. 546–555.
24. He, H.; Shao, Z.; Tan, J. Recognition of Car Makes and Models From a Single Traffic-Camera Image. I*EEE Trans. Intell. Transp. Syst.* **2015**, *16*, 3182–3192. [CrossRef]
25. Das, J.; Shah, M.; Mary, L. Bag of feature approach for vehicle classification in heterogeneous traffic. In Proceedings of the 2017 IEEE International Conference on Signal Processing, Informatics, Communication and Energy Systems (SPICES), Kollam, India, 8–10 August 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 1–5.
26. Pearce, G.; Pears, N. Automatic make and model recognition from frontal images of cars. In Proceedings of the 2011 8th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Klagenfurt, Austria, 30 August–2 September 2011; IEEE: Piscataway, NJ, USA, 2011; pp. 373–378.
27. Gao, Y.; Lee, H.J. Local tiled deep networks for recognition of vehicle make and model. *Sensors* **2016**, *16*, 226. [CrossRef] [PubMed]
28. Elharrouss, O.; Akbari, Y.; Almaadeed, N.; Al-Maadeed, S. Backbones-review: Feature extraction networks for deep learning and deep reinforcement learning approaches. *arXiv* **2022**, arXiv:2206.08016.
29. Lu, L.; Wang, P.; Cao, Y. A novel part-level feature extraction method for fine-grained vehicle recognition. *Pattern Recognit.* **2022**, *131*, 108869. [CrossRef]
30. Alghamdi, A.S.; Saeed, A.; Kamran, M.; Mursi, K.T.; Almukadi, W.S. Vehicle Classification Using Deep Feature Fusion and Genetic Algorithms. *Electronics* **2023**, *12*, 280. [CrossRef]
31. Wu, M.; Zhang, Y.; Zhang, T.; Zhang, W. Background segmentation for vehicle re-identification. In Proceedings of the International Conference on Multimedia Modeling, Daejeon, Republic of Korea, 5–8 January 2020; Springer: Cham, Switzerland, 2020; pp. 88–99.
32. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125
33. Bhatti, H.M.A.; Li, J.; Siddeeq, S.; Rehman, A.; Manzoor, A. Multi-detection and Segmentation of Breast Lesions Based on Mask RCNN-FPN. In Proceedings of the 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Seoul, Republic of Korea, 16–19 December 2020; pp. 2698–2704. [CrossRef]
34. Akbari, Y.; Al-Maadeed, S.; Adam, K. Binarization of degraded document images using convolutional neural networks and wavelet-based multichannel images. *IEEE Access* **2020**, *8*, 153517–153534. [CrossRef]
35. Alom, M.Z.; Taha, T.M.; Yakopcic, C.; Westberg, S.; Sidike, P.; Nasrin, M.S.; Van Esesn, B.C.; Awwal, A.A.S.; Asari, V.K. The history began from AlexNet: A comprehensive survey on deep learning approaches. *arXiv* **2018**, arXiv:1803.01164.
36. LeCun, Y. LeNet-5, Convolutional Neural Networks. Available online: http://yann.lecun.com/exdb/lenet (accessed on 1 December 2022).
37. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.

38. Ballester, P.; Araujo, R.M. On the performance of GoogLeNet and AlexNet applied to sketches. In Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, Phoenix, AZ, USA, 12–17 February 2016.
39. Iandola, F.N.; Han, S.; Moskewicz, M.W.; Ashraf, K.; Dally, W.J.; Keutzer, K. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5 MB model size. *arXiv* **2016**, arXiv:1602.07360.
40. Li, Z.; Liu, F.; Yang, W.; Peng, S.; Zhou, J. A survey of convolutional neural networks: Analysis, applications, and prospects. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, *33*, 6999–7019. [CrossRef]
41. Karlik, B.; Olgac, A.V. Performance analysis of various activation functions in generalized MLP architectures of neural networks. *Int. J. Artif. Intell. Expert Syst.* **2011**, *1*, 111–122.
42. Akbari, Y.; Britto, A.S.; Al-Maadeed, S.; Oliveira, L.S. Binarization of degraded document images using convolutional neural networks based on predicted two-channel images. In Proceedings of the 2019 International Conference on Document Analysis and Recognition (ICDAR), Sydney, NSW, Australia, 20–25 September 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 973–978.
43. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
44. Available online: https://drive.google.com/drive/folders/1zqR1s9YiTxAfjfF213WbiH3Xc-SHPIPs?usp=sharing (accessed on 1 December 2022).
45. Lee, Y.; Park, J. Centermask: Real-time anchor-free instance segmentation. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 13906–13915.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.