Detection of Vehicles in Aerial Photographs Using Convolutional Neural Networks[1*]

Danylo Borovyk[1,†], Roman Fedoniuk[1,†], Sergey Subbotin[1,†], Andrii Oliinyk[1,†] Tetiana Kolpakova[1,†]

[1] *National University "Zaporizhzhia Polytechnic", Zhukovs'koho St, 64, 69063 Zaporizhzhia, Ukraine*

Abstract

Vehicle detection in aerial photography is a crucial step in image processing for many applications such as large area screening. However, compared to ground-based object detection, it remains a challenging task due to the small size of the vehicles and the complex background. Our paper proposes an approach using a double focal loss convolutional neural network (MFL CNN). In this algorithm, we use feedforward communication to improve feature learning in a CNN framework. In addition, the focal loss function replaces the conventional cross- entropy loss function in both the regional proposal network (RPN) and the final classifier.

When developing the algorithm, large-scale data sets of leading scientific companies and universities were used. Featured datasets include EAGLE and XWHEEL. They consist of a large number of aerial photographs of locations with a large number of vehicles, and have a large annotation of classes to identify different vehicles.

By investigating the performance of our model on existing datasets such as XWHEEL and EAGLE, we demonstrate that our MFL outperforms baseline models in vehicle detection.

Keywords

Convolutional Neural network, Object Detection, Focal Loss

1. Introduction

Viewing aerial images covering large areas is critical for many applications such as surveillance, reconnaissance or rescue operations. These applications require accurate identification of all relevant objects, such as vehicles in the camera's field of view, before the scene can be analyzed and interpreted. To reduce the burden on image analysts, a system of automatic object detection is needed.

Typically, vehicle detection in aerial photographs is performed using methods that include manual features and a classifier or a cascade of classifiers within a sliding window approach [1, 2, 3, 4]. Recently, several authors [5, 6, 7] proposed to use convolutional neural networks (CNN) to classify candidate regions. However, calculating convolutional functions for each candidate window separately is computationally expensive [5].

So, methods such as Fast R-CNN [5] and Faster R-CNN [8] showed the most effective results on standard test data sets for detection, significantly reduced the computational time for

SMARTINDUSTRY-2024: International Conference on Smart Automation & Robotics for Future Industry, April 18 - 20, 2024, Lviv, Ukraine

* Corresponding author.

† These authors contributed equally.

romanfedoniuk01@gmail.com (R. Fedoniuk); daniilborovik1999@ukr.net (D. Borovyk) subbotin.csit@gmail.com (S. Subbotin) olejnikaa@gmail.com (A. Oliinyk) t.o.kolpakova@gmail.com (T. Kolpakova)

0000-0001-5814-8268 (S. Subbotin); 0000-0002-6740-6078 (A. Oliinyk); 0000-0001-8307-8134 (T. Kolpakova)

training and testing. Instead of calculating convolution functions separately or using multiple scales, a single convolution feature map is now used for the entire image.

The performance of both methods strongly depends on the so-called object proposal methods, which are used to generate a set of candidate regions as input data for classification. The set of candidate regions should be as limited as possible to reduce the computational effort, while ensuring coverage of all objects. However, both detectors and object proposal methods were developed for datasets that are significantly different from aerial photographs . In general, the images of these datasets, such as Pascal VOC2007 [9], contain only one or a few objects located mainly in the center and occupying a large part of the image. While aerial photographs may include several randomly located objects whose size is in the range of a few pixels.

For a clearer idea, you can view Figure 1.

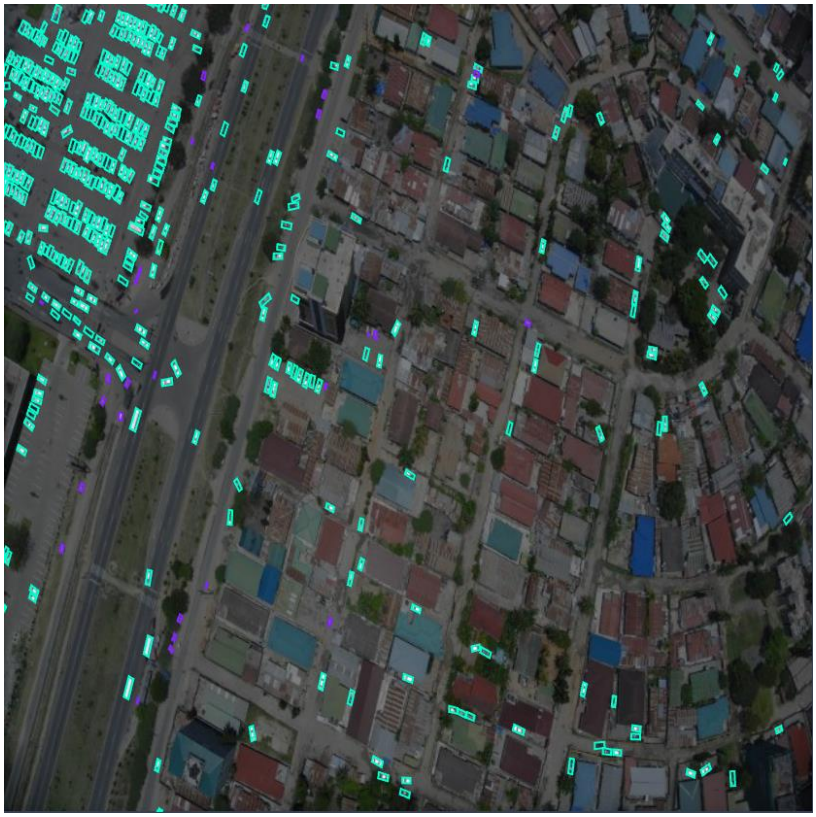


Figure. 1. The image resulting from the annotation of cars (depicted in cyan) and non-cars (represented in purple).

In addition, all these methods localize vehicles due to the use of a sliding window, which leads to significant computational costs. Window sizes and steps of their movement must be carefully adjusted to adapt to different sizes of objects of interest in a given data set [10].

To solve these problems, we developed a special framework for vehicle detection in aerial photographs , shown in Figure 2. This framework, known as a convolutional neural network with dual focal loss (MFL), has three main components:

1. Addition of pass-through coupling from surface to deep layers, allowing for the study of detailed features with a large amount of information.
2. The use of the focal loss function in the regional proposal network (RPN) instead of the standard cross-entropy is aimed at solving the problem of class imbalance [11].

3. Replacing the cross-entropy function with the focal loss function in a classifier to solve the problem of learning on light positive and heavy negative examples.

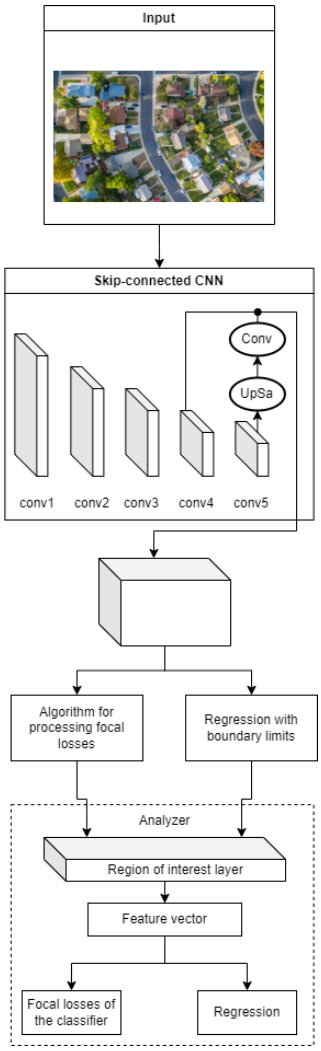


Figure. 2. Summary of the suggested MFL algorithm, comprising three primary components: 1) To study the features, a pass connection from the lower layer to the upper layer is added, which contains rich detailed information. 2) Instead of the traditional cross-entropy, the RPN employs the focal loss function. [11]. 3) The classifier utilizes the focal loss function as a substitute for cross-entropy.

2. Related Work

In the computer vision and photogrammetry literature, object detection and classification are widely studied as fundamental problems. The majority of current methods can traditionally be categorized into three primary stages: first, they decide which areas contain objects of interest, then they extract features and carry out classification. Many of these methods use a sliding window search strategy to generate regions where these objects are likely to be located.

These methods scan images using windows of different scales and locations, which leads to high computational and time complexity, and many of them are inefficient. But Uijlings [12] presented an algorithm known as selective search, which combines the advantages of both

exhaustive search and segmentation. This method is widely used in combination with Deep methods Convolutional Neural Networks (DCNN) for object detection, which made the works of Girshick [13] and [14] famous. In addition, Ren [15] introduced the RPN (Regional Provider Network), which has gained wide acceptance as an approach for generating regional proposals.

Prior to the process of classification, each potential region is identified by features. Kembhavi [16] used scale -invariant feature transform (SIFT) to detect vehicles, Gleason [17] and Han [18] developed methods based on histogram of oriented gradients (HoG), and Bai [19] applied Haar -like signs for this purpose. Despite the effectiveness of their methods, this approach using manually created features is not always effective enough in separating vehicles from complex backgrounds [20, 21, 22, 23]. In recent times, approaches utilizing Deep Convolutional Neural Networks (DCNN) have experienced notable success in the industry of object detection and classification [13, 24, 25, 28].

After obtaining the features, they are submitted to the input of the classifier. The two most widely used classifiers, known for their efficiency and reliability, are stood out by the Random Forest (RF) and Support Vector Machine (SVM). [25]. So far, these have served as the ultimate classifiers in certain CNN-based approaches [13]. Softmax is now the preferred classifier in DCNN-based methods due to its ability to offer normalized probabilistic predictions. Subsequently, cross entropy (CE) is employed to compute losses, which then drive the updating of network parameters [26].

Approaches that have progressed through these three phases are referred to as two-stage methods: the first stage is the proposal of a candidate region, the second stage is object classification. Such two-stage CNN-based methods show the highest results in accuracy. Compared to them, methods that do not require additional operations to offer regions, as follows Single Shot Multibox Detector (SSD) [27] and You Only Look Once (YOLO) [28], are single-stage . They work faster than two-step methods, but at the expense of accuracy. Especially the detection of small-scale objects is a challenge for these approaches. This problem limits their application for vehicle detection in aerial photographs . Therefore, we use a two-step method in our algorithm.

The effectiveness of a method based on deep learning, which has millions of parameters, depends significantly on a large amount of training data. In the past, several large datasets have already been presented for various tasks, such as ImageNet [29] for object classification, Cityscapes for semantic segmentation, etc. [30] consisting of tens of thousands of images for model training. Many of the existing reference datasets contain a variety of vehicles, but they are presented as ground images and are not up to the task of training aerial vehicle detection systems . Some well-annotated datasets for aerial imagery exist , such as VEDAI [31] and XWHEEL. But objects in VEDAI are easily detected due to the sparse number of vehicles and simple background, while XWHEEL, although more complex, is limited to 39 images, of which only 17 (with 8625 vehicles) are used for training. This amount of training data is limited for CNN models.

3. Proposed Algorithm

The description of the proposed algorithm is shown in Figure 2. This algorithm is a modification of the Faster R-CNN standard [15]. For a general object detection procedure, we recommend referring to the work of Ren [15]. In our work, we opted for ResNet [32] as the fundamental framework for feature learning due to its superior efficiency, reliability, and effectiveness in the learning process [33].

3.1. Skip Connection

In the field of object segmentation, it has been determined that features extracted from smaller layers contain more complex details [34]. In a specific scenario of vehicle detection in aerial images, where the size of the vehicle is approximately 30 × 50 pixels, provided that the ground distance (GSD) is 10 cm, the size of the output ResNet object maps after the fifth fusion layer is 32 times smaller than the input size [32]. This reduction in size creates a potential risk of not noticing small vehicles projected onto these maps due to their reduced scale. In addition, the fusion operation at this stage leads to a noticeable loss of detail. In regions characterized by dense traffic, these factors can make it difficult to distinguish individual vehicles. For example, objects derived from shallower layers have more complex details than objects from deeper layers. In densely populated areas, detail becomes a key factor in distinguishing individual vehicles. Therefore, we use an approach that combines features from smaller layers, characterized by greater detail, with features from deeper layers, which provide more representative information. This methodology is depicted in Fig. 3, given an input image size of 748 × 652 pixels. The object maps after the fourth and fifth merging layers have dimensions of 48 × 56 × 1080 and 24 × 28 × 2160, respectfully. To facilitate the fusion, the smaller maps are enlarged to 48 × 56 × 2160 and then reduced to 1080 channels using a 1-to-1 convolution. Subsequently, both feature maps are combined as strip maps.

3.2. Loss Function

The cross-entropy (CE) function is widely used for object classification and reduces the unevenness between positive and negative examples. However, this feature is not efficient enough in distinguishing between simple and complex classification examples, especially in detecting vehicles in aerial photographs . For example, the facades of buildings may look too similar to cars.

The focal loss function has been applied to solve the problem of class imbalance [11] in object detection models such as YOLO [27] and SSD [28]. They showed that single-stage models have a problem with excessive background objects that are difficult to distinguish from vehicles. Two-stage models, such as RPN, solve this problem in the first stage by filtering candidates that are likely to be background, but in complex conditions, such as densely populated areas with cars, this approach is not always effective [15]. A new MFL model was developed, inspired by the concept presented in [9]. This model includes the focal loss function not only at the region proposal stage, but also at the classification stage, solving the problems associated with the complex nature of the task.

The traditional CE loss per classification (for ease of use, we will take binary classification as an example), which is formally defined as:

$$D_{KL}(j, q) = -\log(j_t) \tag{1}$$

where

$$j_t = \begin{cases} j & \text{if } q = 1 \\ 1 - j & \text{in another case} \end{cases}$$

where j is the predicted probability that this candidate will receive a +1 label, and q is the truth label: q ∈ {-1,+1}.

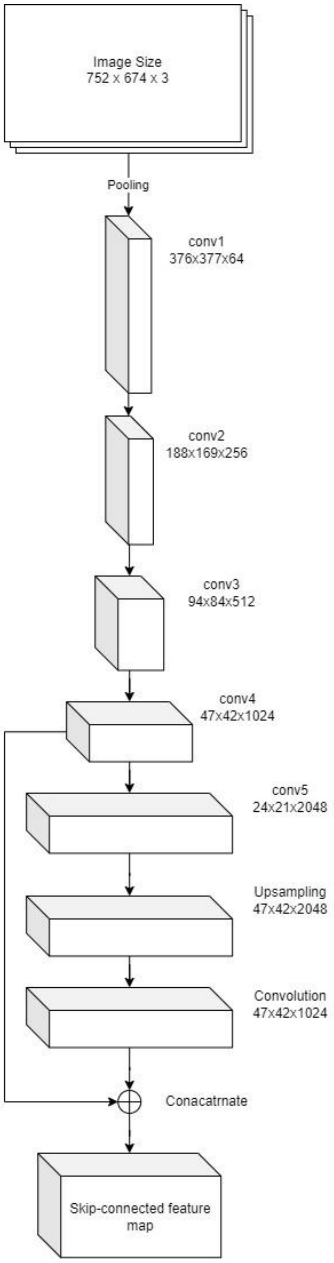


Figure. 3. The gap CNN architecture involves scaling up feature maps from conv5 to sizes corresponding to feature maps from conv4. Next, the number of feature channels is reduced using a 1 × 1 convolution layer to 1024. At the final stage of the map, the attributes of objects from layers 4 and 5 are combined.

Introducing a modulating coefficient, denoted as (1 - jt) ψ, together with an adjustable focusing parameter ψ ≥ 0 into the cross entropy (CE) loss function transforms the loss function into what is called the focus loss (FL):

$$L_{FL}(j_t) = -(1 - j_t)^{\psi} \log(j_t), \quad (2)$$

Focal losses have two key characteristics. First, they have little effect on misclassified examples with low significance (jt) when the modulating coefficient approaches 1. Conversely, as the value of jt increases (jt → 1), the modulating factor approaches 0, which leads to a decrease in losses for correctly classified examples. Second, increasing the focusing parameter (ψ) increases the influence of the modulating factor. The cross-entropy (CE) function can be considered as the partial case at ψ = 0. It is important to note that the contribution of easy examples decreases while the contribution of hard examples increases during the learning process. For example, at ψ = 20 the losses for the example classified with jt = 0.92 are 1% of the CE losses and only 0.1% of them at jt = 0.973.

3.3. Multiple Focal Loss CNN

In our MFL algorithm, we have introduced a pass-through connection that combines features from both the lower (4) and upper (5) layers. This strategic design incorporates a focal loss function in both the Regional Proposition Network (RPN) layer and the eventual classification layer, effectively eliminating class imbalance and solving the problem of distinguishing between easy and difficult examples in our particular task. As mentioned earlier, the final feature maps are reduced by a factor of 16 from the original image size.

To generate candidate proposals, we follow a process in which nine anchor points are generated at the center of each pixel in the object maps. These landmarks span three different scales (90:30, 60:30, 30:30) and three varying scales (9:3, 6:3, and 1:1) based on the initial input images. Each landmark is marked as a true or false example depending on its crossover with a baseline value formally defined using the intersection-over-association (IoA) metric:

$$\mathrm{IoA} = \frac{\mathrm{S(P \cap G)}}{\mathrm{S(P \cup G)}},$$

where the value in the numerator is the area of intersection between the square of the date and the square of the true data, and the value in the denominator is their union. Proposals with an IoA value greater than 0.75 are marked as positive samples, while those with an IoA less than 0.12 are marked as negative. Suggestions that go beyond the image are considered unacceptable. During the training phase, each collection consists of 72 successful and 72 unsuccessful samples.

The loss function for training the Region Proposal Network, which applies focal loss, is calculated using the following formula:

$$L(\{p_i\},\{t_i\}) = \frac{\sum L_{cls-FL}(p_i,p_i^*)}{N_{cls}} + \frac{\lambda \sum_i p_i^* L_{reg}(t_i,t_i^*)}{N_{reg}}, \quad (3)$$

$L_{cls\text{-}FL}$ denotes the focal loss for classification as described in formula 2, L_{reg} denotes the loss for restricted area regression. p_i denotes the expected chance that sentence i belongs to the background, and, while p_i^* denotes its corresponding basic truth tag. N_{cls} represents the sum of the samples, and N_{reg} - represents the sum of the total number of correct samples. The parameter λ is applied to a loss weighting for the restricted regions regression. A plain L1 weighted loss method similar to L_{reg} [15] is used:

$$L_{reg}(v_i, v_i^*) = R(v_i - v_i^*), \quad (4)$$

and

$$R(j) = \begin{cases} 0.5j^2 & if\ |j| < 1 \\ |j| - 0.5 & in\ other\ case \end{cases},$$

$v = (v_x, v_y, v_w, v_h)$ denotes the normalized detailed information about the boundary region for the positive sample, and t* denotes the corresponding basis truth. The formal definition for each of these elements is as follows:

$$\begin{aligned} v_x &= \frac{P_x - A_x}{A_w}, & v_y &= \frac{P_y - A_y}{A_h}, \\ v_w &= \log\frac{P_w}{A_w}, & v_h &= \log\frac{P_h}{A_h}, \\ v_x^* &= \frac{P_x^* - A_x}{A_w}, & v_y^* &= \frac{P_y^* - A_y}{A_h}, \\ v_w^* &= \log\frac{P_w^*}{A_w}, & v_h^* &= \log\frac{P_h^*}{A_h}, \end{aligned} \tag{5}$$

where (Px, Py) represents the coordinates of the center of the intended limit frame, while (Pw, Ph) indicates the intended width and height of the frame. Also, there is information about the binding bounding box $A = (A_x, A_y, A_w, A_h)$. P* denotes information about the truth of the limited frame.

The RPN layer generates a set of candidates that are likely to be objects of interest, such as vehicles in this case, and defines bounding boxes for them. After that, the objects that match these frames are cut from the object maps and passed through the ROI layer to equalize their sizes.

In the final segment of the network, the classifier uses these properties to set labels and make predictions about the constraint frames. The loss function for this subnetwork of the classifier, which relates to each candidate, is formulated as follows:

$$K(P, M) = L_{cls-FL}(P, P^*) + \psi_2 P^* L_{reg}(M, M^*), \tag{6}$$

where M is defined as:

$$\begin{aligned} M_x &= \frac{P_x - A_x}{A_w}, & M_y &= \frac{P_y - A_y}{A_h}, \\ M_w &= \log\frac{P_w}{A_w}, & M_h &= \log\frac{P_h}{A_h}, \\ M_x^* &= \frac{P_x^* - A_x}{A_w}, & M_y^* &= \frac{P_y^* - A_y}{A_h}, \\ M_w^* &= \log\frac{P_w^*}{A_w}, & M_h^* &= \log\frac{P_h^*}{A_h}, \end{aligned} \tag{7}$$

The regions of predictions, anchors, and basic truths are denoted by P_x, A_x, and P_x*, respectively, and the sub-indices y, w, and h fulfill similar functions. The parameter ψ_2 is equal to 1 to ensure that both classification and limiting frames are equally affected. In the process of training, the subnetwork of the classifier is trained in a proportion of 1:3 for successful and unsuccessful samples, following the standard training methodology [15].

3.4. EAGLE Dataset

In this topic, we will look at a dataset that was used for training called EAGLE. The dataset is used to detect vehicles of various types, including determination of vehicle direction using aerial images.

The dataset comprises a collection of high-quality air photos reflecting a variety of real-world scenarios, including variations in imaging detectors, angles, hours, heights, density (from 5 to 45 cm in pixels on the terrain), climate and illumination conditions, and city and villages [35]. The data for the set was acquired in the period from 2006 to 2019. EAGLE contains 215,986 annotated vehicles in 318 aerial images covering both small vehicles (such as police cars, ambulances, passenger cars, transporters, minivans, and off-road vehicles) and large vehicles (including vans, trucks, buses, heavy trucks, construction vehicles, fire trucks, and trailers). The annotations include orientation boxes marked with four points [35]. The pictures are presented in the form of files with the JPG extension, the size of which is 5616 x 3744 pixels, and the annotation file is presented in XML format. The annotation contains the corresponding coordinates of all four corners of the vehicle, as well as the degree of orientation from 0° to 360°, which indicates the angle of inclination of the vehicle. In addition, for each example, the clarity (completely/partially/poorly) and the ability to determine the orientation of the vehicle (clear/unclear) are indicated [35].

4. Experiments

In this topic, we present the experimental setup and subsets of data used to implement the suggested method and compare it with the most advanced object detectors.

4.1. Dataset and Experimental Settings

Our approach is evaluated on the EAGLE and XWHEEL datasets. Table 1 presents statistical information about these data sets. Both sets use the state-of-the-art Faster R-CNN object detector, which creates a robust baseline for the input data.

Table 4.1
Statistics of EAGLE dataset and XWHEEL dataset.

	Training Set	**Testing Set**	**Image Size**
EAGLE	159 images (108,215 vehicles)	106 images (70 433 vehicles)	5475 × 3345
XWHEEL	39 images (8704 vehicles)	17 images (3347 vehicles)	5472 × 3456

To make efficient use of GPU memory, each initial image from the dataset is split into small pieces of the same size. The resulting fragments have a resolution of 376 × 377 pixels. The position information in the annotations is adjusted to match the corresponding truncated areas. An XWHEEL dataset annotates every transport object by identifying a box that closely fits it. To set up our experiment, we converted the initial annotations into regular square frames defined by a midpoint, altitude, and wide.

Keras implementations of deep learning models use TensorFlow as their backend. [36]. The ResNet-50 network serves as the foundational architecture for feature learning in both Faster R-CNN and our model. For RPN training, we use a training rate of 0.00001. It should also be noted that our algorithm can use other CNN architectures, such as VGGnet [37] or Google

Incejtion [38]. It is important to note that CNN structures are pre-trained on the ImageNet dataset [39].

To evaluate the results of the experiment, the accuracy and F1-balance metrics are used, which are officially defined as follows:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{8}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{9}$$

$$\text{F1} = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}}. \tag{10}$$

TP, FN, and FP denote respectively truthfully correct, negatively correct, and positively valid results. In addition, the connections among IoA and recall, accuracy, and precision, respectively, are also discussed.

4.2. Results on EAGLE Dataset

We evaluated our MFL method on the demanding EAGLE dataset. We used the state-of-the-art Faster R-CNN object detector to create a reliable base estimate [15]. In addition, as a weak baseline estimator, we used the traditional HOG + SVM method [40].

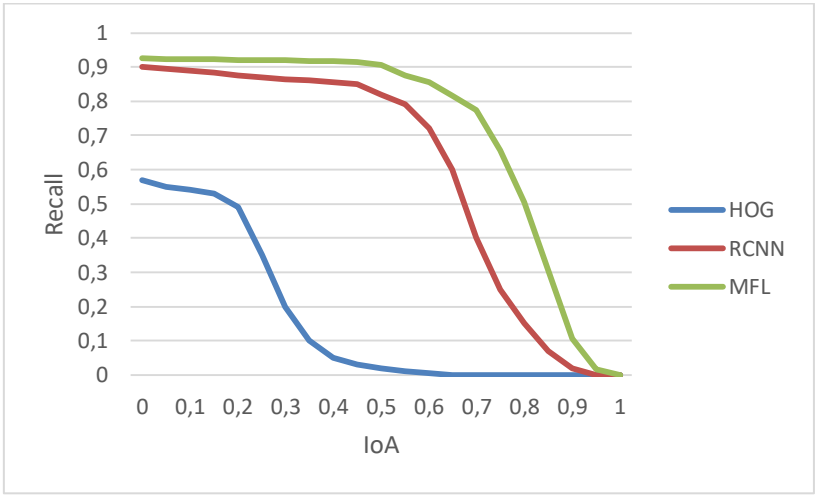
Figure 4 shows how the rate of correct identification varies compared to the accuracy of the three algorithms: MFL, Faster R-CNN, and HOG + SVM, using different IoA values on the EAGLE dataset. It is obvious that methods based on deep learning (MFL - green line and Faster R-CNN - red line) significantly outperform the traditional method (HOG + SVM - black line). Regarding the ratio between recall and precision, our MFL method is found to be more efficient than Faster R-CNN. These curves show that IoA = 0.3 is the optimal balance point for future experiments, providing high speed and accuracy at the same time - a commonly accepted value for object detection tasks. The experimental output of these approaches is presented in Table 2 (at IoA = 0.3), where it is shown that our method performs better than the others.

Table 4.2
Analysis of basis lines and MFL model in EAGLE.

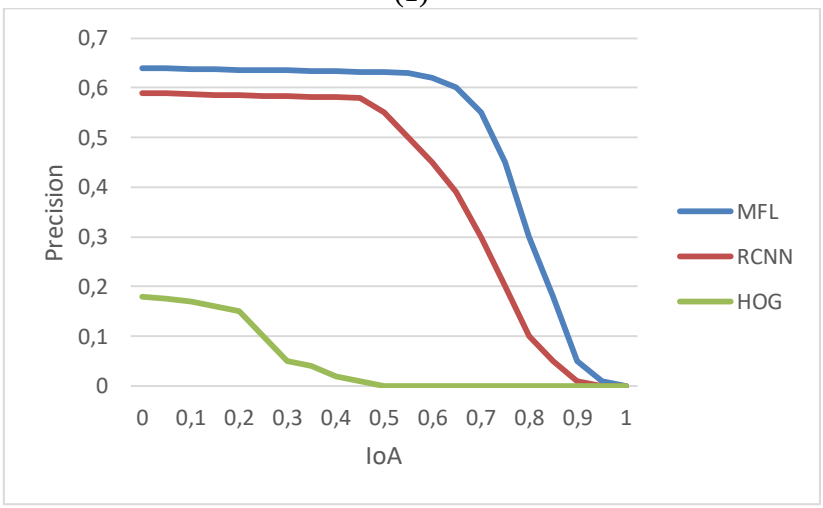
	HOG	**RCNN**	**MFL**
Recall	24.95%	88.28%	89.07%
Precision	11.84%	63.58%	68.79%
F1	0.1606	0.7392	0.7763

We have performed extensive ablation studies to demonstrate the benefits of using pass-through coupling and focal loss functions. Initially, two frameworks were trained that used a dual focal loss function. One of them had a connection of character cards with a pass, and the other did not. The results are shown in Figure 5. We noticed that boundary field predictions made using the cross-linked feature map structure were significantly more accurate than in the case where there is no cross-connection. Also, individual vehicles were better distinguished from others thanks to the use of small signs. Then, two additional feedback structures were trained, one using the CE loss function and the second one using the dual loss function. The quality output is shown in Fig. 6. The framework trained using the CE feature showed a tendency to misidentify many background objects that look similar to vehicles as real vehicles.

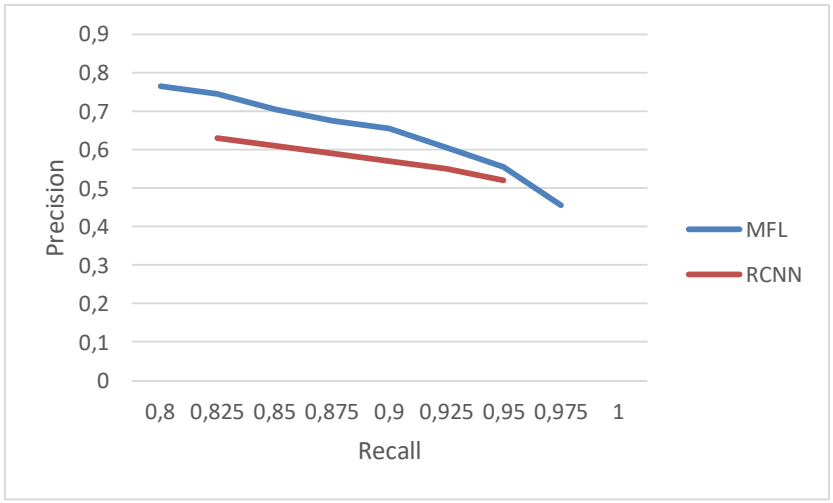
At the same time, the algorithm using double focal loss for training proved to be significantly more effective in distinguishing these complex negative examples from real vehicles.



(1)



(2)



(3)

Figure. 4. Correlation between IoA and recall rate (1), IoA and accuracy (2), and recall rate and accuracy (3) for MFL, Faster R-CNN, and HOG + SVM models on the EAGLE dataset.

(1) (2)

Figure. 5 . A comparative analysis of region boundary prediction for a network without a gap (1) and a network with a gap (2) shows differences in the quality of prediction. It is worth noting that the frames provided by the network with the connection of object map skips are significantly more accurate than those created by the network without connection (highlighted in yellow). Other parameters remain unchanged.

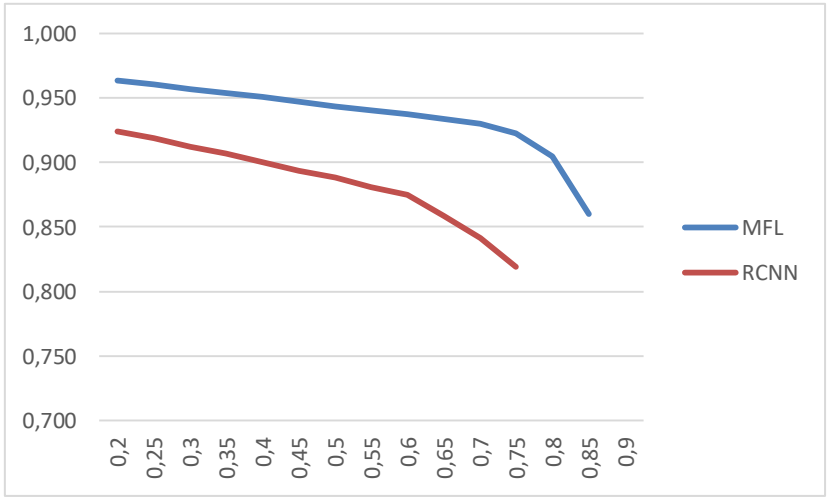
(1) (2)

Figure. 6. Comparison of the efficiency of car detection by different frameworks trained by two different loss functions - CE loss (a) and FL (b). Other parameters remain the same.

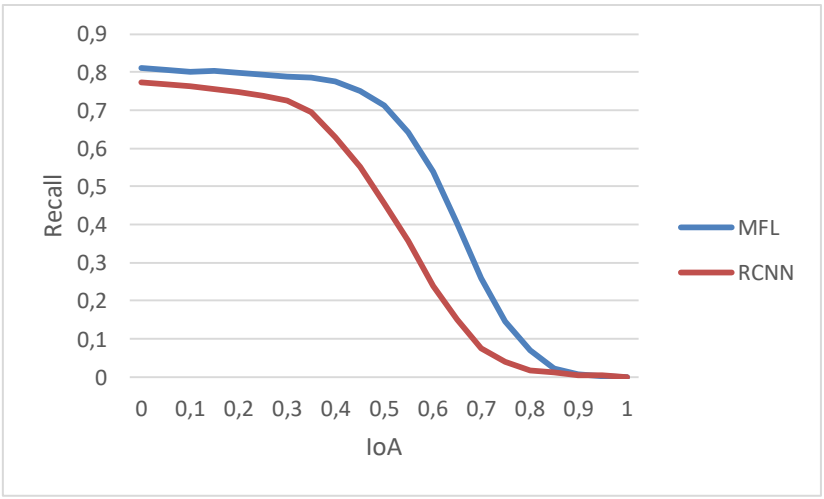
(1) (2)

Figure. 7. Examples of false detection by our model are shown in red thin line, displaying the detection results. Cars that were missed are marked in green, while false positives are in blue.

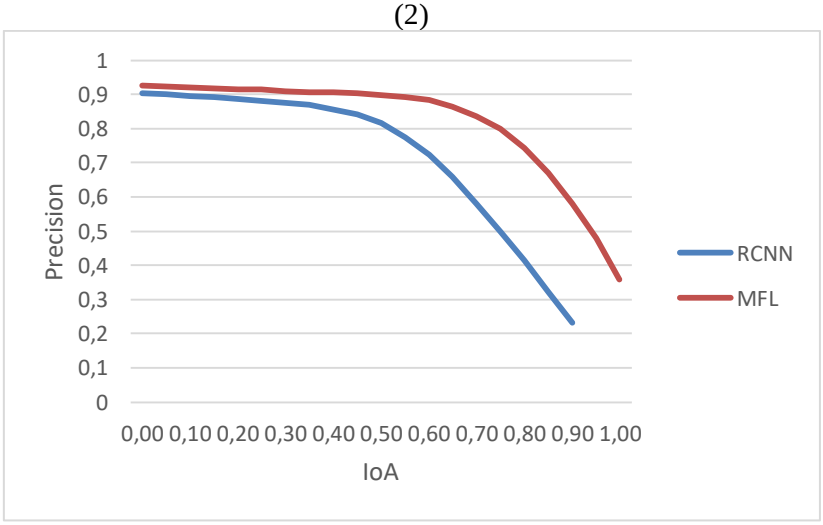
Figure 7 shows examples of shortcomings in the work of the proposed detection method. Although our detection approach showed significant improvements in accuracy and detection rate compared to the baseline methods, some vehicles still remain undetected, especially in quiet parking lots depicted in Figure 7(1). Otherwise, certain items that look like vehicles are falsely identified as vehicles, as shown in Figure 7(2).



(1)



(2)



(3)

Figure. 8. Correlation between response and accuracy (a), IoA and accuracy (b), and IoA and response rate (c) for MFL and Faster R-CNN on the XWHEEL dataset.

(1) (2)

Figure. 9. Comparison of the detection quality between the results of Faster R-CNN (a) and MFL (b) on the XWHEEL dataset.

Our model was tested on the XWHEEL dataset as well. Fig. 8 shows the correlation between recall speed and accuracy for both RCNN and the proposed method. Furthermore, Fig. 8

emphasizes that our method outperforms the benchmark RCNN in both recall speed and accuracy.

To further evaluate the performance, we evaluated RCNN and MFL, especially in scenarios with densely parked cars in the XWHEEL dataset, as shown in Figure 9. Qualitative results show that MFL (Figure 9(2)) discovered more detached vehicles and provided more accurate constraint frames than RCNN (Figure 9(1)).

To prove the effectiveness of our approach for detecting objects in air photos, we conducted a comparative analysis of our experimental results with some other methods such as Hyper Region Proposal Network [41], Fast Multiclass Vehicle Detection [3] and Shallow YOLO [21] on the XWHEEL dataset. The outcomes of these comparisons are presented in Table 3. Our method significantly outperforms FMVD and Shallow YOLO in all three metrics. Compared to HRPN, our approach shows only a small superiority (2% for F1). However, it is worth noting that HRPN uses a cascade of classifiers improved by extracting negative examples. This likely results in increased computational overhead and may cause a class imbalance problem. Our method operates with the focal loss function, so it avoids such problems.

Table 3
Comparison of experimental results between FMVD, Shallow YOLO, HRPN and our approach on the XWHEEL dataset.

	FMVD	Shallow	HRPN	MFL
Recall	66.91%	65.8%	75.82%	78.53%
Precision	82.36%	54.21%	88.39%	89.56%
F1	0.7383	0.5944	0.8163	0.8368

5. Conclusion

In this paper, we deployed a customized MFL architecture for the purpose of vehicle detection in aerial imagery. Our approach combines feature properties from the lower and upper layers of the network to improve the ability to distinguish individual vehicles in crowded scenes. To solve the problems associated with class imbalance and example complexity, we chose the focal loss function in both the feature region suggestion and classification phases instead of using cross-entropy. During training, we used the large EAGLE dataset, which includes annotations for all vehicles in the scene, covering a large number of objects. Experimental results demonstrate the superiority of our method over the classical ones on two datasets. In the future, we plan to extend MFL to recognize vehicle types and determine their orientation.

References

[1] Oliinyk , A., Fedorchenko , I., Stepanenko, A., Katschan , A., Fedorchenko , Y., Kharchenko, A., Goncharenko, D. "Development of genetic methods for predicting the incidence of volumes of emissions of pollutants in the air". 2019 2nd International Workshop on Informatics and Data-Driven Medicine, IDDM, CEUR Workshop Proceedings, 2019, Vol.2488, pp. 340–353.

[2] Cheng, G., Han, J. A survey on object detection in optical remote sensing images. ISPRS Journal of Photogrammetry and Remote Sensing, 117:11–28 (2016).

[3] Liu K., Mattyus G. Fast multiclass vehicle detection on aerial images. Geoscience and Remote Sensing Letters, IEEE, PP(99):1–5 (2015).

[4] Moranduzzo , T., Melgani , F. Detecting cars in UAV images with a catalog-based approach. IEEE Transactions on Geoscience and Remote Sensing, 52(10):6356–6367 (2014).
[5] Chen X., Xiang S., Liu C., Pan C. Vehicle detection in satellite images by hybrid deep convolutional neural networks. IEEE Geoscience and Remote Sensing Letters, 11(10):1797–1801 (2014).
[6] Girshick R., Donahue J., Darrell T., Malik J. Rich feature hierarchies for accurate object detection and semantic segmentation. Proceedings IEEE Conference on Computer Vision and Pattern Recognition, pages 580–587 (2014).
[7] Fedorchenko , I., Oliinyk , A., Stepanenko, A., Zaiko , T., Korniienko , S., Burtsev , N. "Development of a genetic algorithm for placing power supply sources in a distributed electric network". European Journal of Enterprise Technologies, issue 5/101, 6–16 (2019), doi : 10.15587/1729-4061.2019.180897
[8] Everingham M., Van Gool L., Williams C., Winn J., A. Zisserman. The Pascal Visual Object Classes (VOC) challenge. International Journal of Computer Vision, 88(2):303–338 (2010).
[9] Shkarupylo V., Kudermetov R., Timenko A., Polska O. On the Aspects of IoT Protocols Specification and Verification. Problems of Infocommunications. Science and Technology : 2019 International Scientific-Practical Conference, PIC S&T′2019 (Kyiv, Ukraine, October 8-11, 2019). P. 93-96. DOI: https://doi.org/10.1109/PICST47496.2019.9061406
[10] Fedorchenko , I., Oliinyk , A., Stepanenko, A., Svyrydenko , A, Goncharenko, D. "Genetic method of image processing for motor vehicle recognition". 2019 2nd International Workshop on Computer Modeling and Intelligent Systems, CMIS, 2019, Zaporizhzhia, April 15-19, CEUR Workshop Proceedings, Vol. 2353, pp. 211-226.
[11] Lin, T., Goyal, P., Girshick , RB, He, K., Dollár , P. Focal loss for dense object detection. Proceedings International Conference on Computer Vision. pages 2999–3007 (2017).
[12] Shkarupylo V. V., Tomičić I., Kasian K. M. The investigation of TLC model checker properties. Journal of Information and Organizational Sciences, 2016. Vol. 40, No. 1. P. 145-152. DOI: https://doi.org/10.31341/jios.40.1.7
[13] Girshick , R., Donahue, J., Darrell, T., Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. Proceedings IEEE Conference on Computer Vision and Pattern Recognition. pages 580–587 (2014).
[14] Girshick , R. Fast R-CNN. In Proceedings of the IEEE International Conference on Computer Vision, pages 1440–1448 (2015).
[15] Alsayaydeh J. A. J., Indra W. A., Khang A. W. Y., Zakir Hossain A. K. M., Shkarupylo V., Pusppanathan J. The experimental studies of the automatic control methods of magnetic separators performance by magnetic product. ARPN Journal of Engineering and Applied Sciences, April 2020. Vol. 15, No. 7. P. 922–927. DOI: https://doi.org/10.5281/zenodo.5163618
[16] Kembhavi , A., Harwood, D., Davis, LS Vehicle detection using partial least squares. IEEE Transactions on Pattern Analysis and Machine Intelligence, 33(6):1250–1265 (2011).
[17] Shkarupylo V., Blinov I., Chemeris A., Dusheba V., Alsayaydeh J., Oliinyk A. Iterative Approach to TLC Model Checker Application. Proc. 2021 IEEE KhPI Week on Advanced Technology (Kharkiv, Ukraine, September 13 – 17, 2021). P. 283–287. DOI: https://doi.org/10.1109/KhPIWeek53812.2021.9570055

[18] Han, F., Shan, Y., Cekander, R., Sawhney, HS, Kumar, R. A two-stage approach to people and vehicle detection with HOG based SVM. Proceedings Performance Metrics for Intelligent Systems Workshop. pages 133–140 (2006).
[19] Bai, H., Wu, J., Liu, C. Motion and Haar-like features-based vehicle detection. In Proceedings International Conference on Multi-Media Modeling (2006).
[20] Krizhevsky , A., Sutskever , I., Hinton, GE Imagenet classification with deep convolutional neural networks. Proceedings Advances in Neural Information Processing Systems. pages 1097–1105 (2012).
[21] Tang, T., Zhou, S., Deng, Z., Zou, H., Lei, L. Vehicle detection in aerial images based on region convolutional neural networks and hard negative example mining. Sensors, 17(2):336 (2017).
[22] Oliinyk , A., Fedorchenko , I., Stepanenko , .Rud M., Goncharenko, D. Implementation of evolutionary methods of solving the traveling salesman problem in a robotic warehouse // Lecture Notes on Data Engineering and Communications Technologies, 2021, 48, P. 263–292.
[23] Fedorchenko , I., Oliinyk , A., Stepanenko, Zaiko , T., Korniienko S., Kharchenko, A. Construction of a genetic method to forecast the population health indicators based on neural network models // Eastern-European Journal of Enterprise Technologies, 2020, 1 (4-103), P. 52–63. DOI: 10.15587/1729-4061.2020.197319
[24] Carlet, J., Abayowa , B. Fast vehicle detection in aerial imagery. arXiv preprint arXiv:1709.08666 (2017).
[25] Zhao, T., Nevatia , R. Car detection in low-resolution aerial images. Image and Vision Computing, 21(8):693–703 (2003).
[26] LeCun, Y., Bengio, Y., Hinton, G. Deep learning. Nature, 521(7553):436–444 (2015).
[27] Shkarupylo, V.V., Blinov, I.V., Chemeris, A.A., Dusheba, V.V., Alsayaydeh, J.A.J., 2021. On Applicability of Model Checking Technique in Power Systems and Electric Power Industry. Studies in Systems, Decision and Control, book series (SSDC, volume 399), pp. 3–21. (SCOPUS).
[28] Indra, W.A., Zamzam, N.S., Saptari, A., Alsayaydeh, J.A.J, Hassim, N.B., 2020." Development of Security System Using Motion Sensor Powered by RF Energy Harvesting", 2020 IEEE Student Conference on Research and Development, SCOReD 2020 9250984, pp. 254-258.
[29] Nurul Fazleen Binti Abdul Rahim, Adam Wong Yoon Khang, Aslinda Hassan, Shamsul Jamel Elias, Johar Akbar Mohamat Gani, Jamaluddin Jasmis, Jamil Abedalrahim Jamil Alsayaydeh, "Channel Congestion Control in VANET for Safety and Non-Safety Communication: A Review," 2021 6th IEEE International Conference on Recent Advances and Innovations in Engineering (ICRAIE), 2021, pp. 1-6.
[30] Cordts M., Omran M., Ramos S., Rehfeld T., Enzweiler M., Benenson R., Franke U., Roth S., Schiele B. The Cityscapes dataset for semantic urban scene understanding. In Proceedings IEEE Conference on Computer Vision and Pattern Recognition (2016).
[31] Rezakarivony S., Jurie F. Vehicle detection in aerial imagery: A small target detection benchmark. Journal of Visual Communication and Image Representation, 34:187–203 (2016).

[32] 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. pages 770–778 (2016).
[33] Canziani A., Paszke A., Culurciello E. An analysis of deep neural network models for practical applications. arXiv preprint arXiv:1605.07678 (2016).
[34] Win Adiyansyah Indra, Adam Wong Yoon Khang, Yap Thai Yung and Jamil Abedalrahim Jamil Alsayaydeh, 2019. Radio-Frequency Identification (RFID) Item Finder Using Radio Frequency Energy Harvesting. ARPN Journal of Engineering and Applied Sciences. (VOL. 14 NO. 20) (P. 3554-3560).
[35] Azimi S, Bahmanyar R., Henry C., Kurz F., "EAGLE: Large-scale Vehicle Detection Dataset in Real-World Scenarios using Aerial Imagery," in International Conference on Pattern Recognition (ICPR), 2020. https ://ieeexplore.ieee.org/document/9412353, last accessed 2024/01/15
[36] Abadi M., Agarwal A., Barham P., Brevdo E., Chen Z., Citro C., Corrado GS, Davis A., Dean J., Devin M. Tensorflow : Large scale machine learning on heterogeneous distributed systems. arXiv preprint arXiv:1603.04467 (2016).
[37] Zakir Hossain, A. K. M., Hassim, N. B., Alsayaydeh, J. A. J., Hasan, M. K., & Islam, M. R. (2021). A tree-profile shape ultra wide band antenna for chipless RFID tags. International Journal of Advanced Computer Science and Applications, 12(4), 546-550. doi:10.14569/IJACSA.2021.0120469.
[38] Szegedy C., Vanhoucke V., Ioffe S., Shlens J., Wojna Z. Rethinking the inception architecture for computer vision. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pages 2818–2826 (2016).
[39] Fedorchenko I., Oliinyk A., Jamil Abedalrahim Jamil Alsayaydeh*, Kharchenko A., Stepanenko A., and Shkarupylo V.2020. MODIFIED GENETIC ALGORITHM TO DETERMINE THE LOCATION OF THE DISTRIBUTION POWER SUPPLY NETWORKS IN THE CITY. ARPN Journal of Engineering and Applied Sciences. (VOL. 15 NO. 23) (pp 2850-2867).
[40] Adam Wong Yoon Khang, Shamsul J. Elias, Nadiatulhuda Zulkifli, Win Adiyansyah Indra, Jamil Abedalrahim Jamil Alsayaydeh, Zahariah Manap, Johar Akbar Mohamat Gani, 2020. Qualitative Based QoS Performance Study Using Hybrid ACO and PSO Algorithm Routing in MANET. Journal of Physics, Conference Series 1502 (2020) 012004, doi:10.1088/1742-6596/1502/1/012004.
[41] Jamil Abedalrahim Jamil Alsayaydeh*, Azwan Aziz, A. I. A. Rahman, Syed Najib Syed Salim, Maslan Zainon, Zikri Abadi Baharudin, Muhammad Inam Abbasi and Adam Wong Yoon Khang, 2021. DEVELOPMENT OF PROGRAMMABLE HOME SECURITY USING GSM SYSTEM FOR EARLY PREVENTION, ARPN Journal of Engineering and Applied Sciences. (VOL. 16 NO. 1) pp 88-97.