Can Tho University Journal of Science

website: ctujs.ctu.edu.vn

DOI: 10.22144/ctu.jen.2023.006

Object detection by the combination of generic roi extractor and dynamic R-CNN with side-aware boundary localization in aerial images

Bao Tran Nguyen*, Tai Pham Tan, Doanh C. Bui, Nguyen D.Vo and Khang Nguyen

University of Information Technology, Ho Chi Minh, Viet Nam

**Correspondence: Bao Tran Nguyen (email: 20520142@gm.uit.edu.vn)*

Article info.

Received 21 Apr 2022
Revised 27 May 2022
Accepted 08 Jun 2022

Keywords

Object detection, UAV datasets, vehicle detection, XDUAV dataset

ABSTRACT

Unmanned Aerial Vehicles (UAVs) have recently gained popularity due to their simplicity and effectiveness in traffic monitoring and potential for rapid delivery, and rescue support. Moreover, UAVs have been employed as a supporting machine in data collection for object detection tasks, in particular vehicle detection tasks in object recognition. Although vehicle identification is a tough problem, many of its challenges have recently been overcome by two-stage approaches such as Faster R-CNN, one of the most successful vehicle detectors. However, many critical problems still remain, such as partial occlusion, object truncation, object multi-angle rotation, etc. In this paper, we combine the Generic RoI Extractor (GroIE) method with Dynamic R-CNN and Side-aware Boundary Localization (SABL) for both testing and evaluation on a challenging dataset XDUAV. Overall, 4344 images in the XDUAV dataset, divided into 3 subsets: 3485 training images, 869 testing images and 869 validating images were used. These consisted of six object classes: 33841 "car"; 2690 "bus"; 2848 "truck"; 173 "tanker"; 6656 "motor" and 2024 "bicycle". With the ResNet-101 backbone, our approach showed competitive results compared with the original GRoIE method, surpassed by 1.2% on mAP score and by about 2% on most classes AP scores, except for the class 'tanker'.

1. INTRODUCTION

The introduction and potential of Unmanned Aerial Vehicles (UAVs) have sparked a surge in the object detection research field, particularly in aerial image detection. There are a large volume images or videos captured by drone detection supplied as UAV datasets every year, which poses numerous challenges for object detection. Some of the most crucial problems that can be listed are unrecognizable objects due to trees or building obstruction, confusion between objects, or the unbalanced number of instances in different classes. These difficulties decrease the quality of object detectors significantly.

However, recent studies that have addressed drone detection enhance the "vision" of the computer to recognize and detect objects effectively. For two-stage methods, one of the approaches to improve the robustness and accuracy of detectors is to increase the quality of Regions of Interest (RoI). Moreover, most drone detection problems are instance segmentation, where the output is a set of rectangular bounding boxes representing the localization and classification of each object sample. In addition to RoI improvement, the improvement of the bounding box regressor also helps overcome the challenges of UAV detection. Therefore, many state-of-the-art modules propose enhanced RoI modules or enhanced bounding-box schemes, such as Cascade R-CNN (Cai et al., 2018), and VistrongerDET (Junfeng et al., 2021).

This paper investigates and evaluates the combination of applying GRoIE method, Dynamic R-CNN, and Side-aware Boundary Localization with different backbones on the XDUAV dataset.

2. RELATED WORKS

Two-stage method: With the huge accomplishments of the R-CNN family, two-stage modules have grown in popularity. Object detection and object recognition are the two fundamental responsibilities of two-stage detectors. Specifically, R-CNN (Girshick et al., 2014) module proposes regions of interest within the Selective Search scheme at the beginning, and thereafter the support vector machine (SVM) classifies these proposal regions. After R-CNN, the author of the R-CNN method continued to develop the Fast R-CNN (Girshickt et al., 2015) inspired by Spatial Pyramid Pooling Network (SPPNet) (He et al., 2015). Instead of the Selective Search at the beginning, Fast R-CNN integrated a convolution network for all input images. Although the idea of Fast R-CNN is based on SPPNet, this module differs from the multi-level spatial pyramid pooling layer-SPPNet in utilizing the RoI Pooling as a single-level SPP layer. In addition, the bounding box regressor within Fast R-CNN was adjusted according to the outputs of the softmax and linear layers, which improved the model's speed and surpassed the SPPNet. Continuing the great success of R-CNN and Fast R-CNN, Faster R-CNN (Ren et al., 2015) continues to show outstanding results in object detection. Faster R-CNN stands out better than its predecessor with the Region Proposal Network (RPN). RPN produces regions of interest from feature maps convolved by a convolution network. The output of RPN is divided into two components: binary object classification, which classifies objects from the background and bounding box regression, which determines the confident regions of objects; therefore, RPN comprises two loss functions.

Moreover, for the image segmentation task, an extension of Faster R-CNN called Mask R-CNN (He et al., 2017) was proposed. The overview of image segmentation is that a module will split digital images into many image segments, which can reduce image information to something more relevant and easier to study. In addition, there are two kinds of image segmentation: semantic segmentation and instance segmentation. Specifically, in semantic segmentation, each pixel is distributed into a set. In fact, semantic segmentation determines or classifies objects into only one class at the pixel level. For instance, segmentation is the object determination that separates explicit objects. In conclusion, Mask R-CNN is introduced to solve image segmentation tasks with object masks as output features.

Backbones: Besides the significant improvement of the two-stage module, backbone architecture received enormous upgrade from the basic convolutional network. To have deep training with the convolutional neural network, a residual network (ResNet) (He et al., 2016) was built to solve this problem. Before ResNet, there was a challenging problem in deep learning, called vanishing gradient. With the help of “Skip Connections” between residual blocks, ResNet developed a successful deep training model. Although ResNet brought many colorful results in object detection and became one of the most effective backbones for models, for a variety of different datasets, there were still a large number of hyper-parameters that need adjusting. Thus, ResNeXt was introduced by Xie et al. (2017) as an upgrade to ResNet, which reduced the number of required hyperparameters from ResNet thanks to the new dimension called “cardinality”. Cardinality presents the complexity of transformations. In addition, another approach based on ResNet entitled ResNeSt (Zhang et al., 2020b) was demonstrated to extract more information about cross-channel that ResNet could not achieve. Inspired by ResNeXt, ResNeSt also integrated cardinality within its structure. The difference was the combination of Attention of Squeez and Excitation Net together to formulate a Split-Attention model, which helped boost the interdependencies of channels without additional computational cost.

3. METHODOLOGY

3.1. Experimental Object Detection Methods

3.1.1. GRoIE

Generic RoI Extractor (Rossi et al., 2021), also called GRoIE, is an improvement on the existing RoI extractors that use only one (the best) layer from FPN. The GRoIE approach overcomes the constraint of standard RoI extractors by leveraging all FPN layers since PANet is intuitively based research with each layer retaining valuable information (Lin et al.,2018) Additionally, the GRoIE method can be applied in every two-stage architecture, which performs object detection or instance segmentation.

There are four primary parts that correspond to the four modules: RoI pooling, pre-processing, aggregation, and post-processing. The RoI pooler module, RoI Align (He et al.,2017), is utilized in the first module to pool from the region formed by the RPN. The target of this module is to obtain a fixed-size RoI. The pre-processing module and the aggregating module then turn these features into a single FPN by summation. Finally, the post-processing module is utilized to obtain global features and omit redundant information.

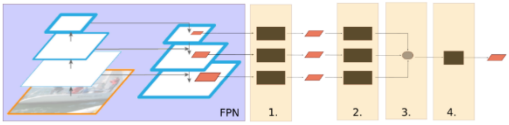


Figure 1. Generic RoI Extraction framework

(1) RoI Pooler. (2) Preprocessing. (3) Aggregation function. (4) Post-processing (Rossi et al., 2021)

3.1.2. Dynamic R-CNN

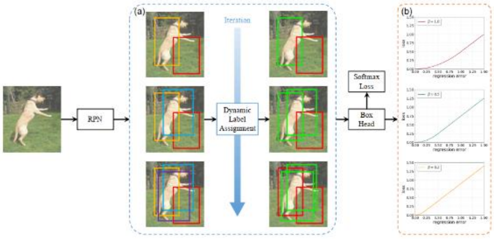


Figure 2. The overall pipeline of the proposed Dynamic R-CNN

(Zhang et al., 2020a)

Zhang et al. (2020a) introduced Dynamic R-CNN as a two-stage detector which helps achieve high-quality object detection. There are two main enhancements in Dynamic RCNN compared to its predecessor: Dynamic Label Assignment (DLA) and Dynamic SmoothL1 Loss (DSL). Based on the standard proposal classification method in Faster R-CNN, Dynamic Label Assignment deploys a dynamic IoU threshold instead, which can be formulated as follows:

$$label = \begin{cases} 1, if \max IoU(b,G) \geq T_{now} \\ 0, if \max IoU(b,G) < T_{now} \end{cases}$$

The Dynamic IoU threshold allows for teh gradual increase in the IoU threshold to achieve the highest possible IoU threshold, which lets the module achieve a high-quality object detection level. In addition, increasing IoU directly at the beginning can vanish positive samples. Another enhancement of Dynamic R-CNN is in the SmoothL1 Loss, the authors proposed to use hyper-parameter β in a changeable way, which means that β will be transformed to β_{now} which is adjusted automatically in every iteration just like the updated mechanism of T_{now}. However, the difference between the adoption of β_{now} and T_{now} is that β_{now} receives the median value in each batch instead of the mean value one like T_{now}.

3.1.3. Side-Aware Boundary Localization

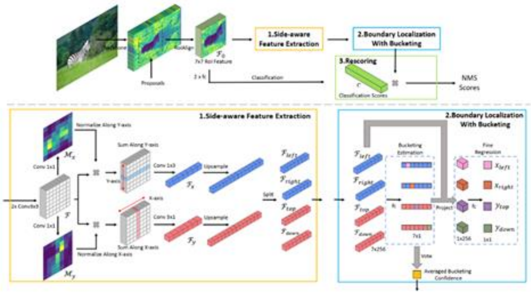


Figure 3. The pipeline of Side-Aware Boundary Localization (SABL) for the two-stage detector

(Wang et al.,2020)

Side-aware Boundary Localization (SABL) (Wang et al., 2020) is a method that is an enhancement on the standard bounding box regression branch to resolve the existing displacements with large variance problems. SABL can be integrated into every one-stage or two-stage detection model. Moreover, in SABL, (Wang et al., (2020proposed a new object localization scheme that has three main parts in the framework: side-aware feature extraction, boundary localization with bucketing and bucketing-guided rescoring.

To extract side-aware features, from the features of $k \times k$ RoI maps, providing side-aware features as $F_{left}, F_{right}, F_{top}$ and F_{down}. These features are utilized to obtain boundaries at each side of the bounding box, which are nearest the ground-truth scale respectively. In the final step, rescoring these bounding boxes occurs, to obtain the bounding boxes that have high classification confidence and accurate localization.

3.2. Experimental Loss Functions

3.2.1. Cross-Entropy Loss

Overall, Cross-Entropy loss, which is based on the theory of entropy, is a metric or a loss function used for evaluating the robustness of the classification model. Specifically, Cross-Entropy loss defines the distance of differences between two probability distributions of the classification model and the predicted distribution, the lower of cross-entropy loss, the better the model demonstrates. The Cross-Entropy loss can be demonstrated in the formula as:

$$L_{CE} = -\sum_{i=1}^{n} y_i \ln(p_i)$$

where n stands for the number of classes, y_i is the label (1 for object, 0 otherwise) and p_i is the probability of the i^{th} class.

3.2.2. Focal Loss

Focal Loss (LIN et al.,2020) is a loss function improved from Cross-Entropy Loss (CE) that assigns more weights to difficult objects or misclassified samples and reduces the weight of easy examples to reduce the risk of the imbalance problem. The formula of focal-loss function is demonstrated as:

$$L_{FL} = -\sum_{i}^{n} y_i (1 - p_i)^{\gamma} \ln(p_i)$$

The difference between Focal-Loss and Cross-Entropy loss is the appearance of $(1 - p_i)^{\gamma}$, hich affects both loss function and gradient descent.

With $(1 - p_i)^{\gamma}$, the easily classified examples, which usually has p_i close to 1, will have less influence on the loss, while the focus on difficult samples is improved.

3.3. The hybrid approach of combining GroIE with Dynamic R-CNN and SABL

To enhance the evaluation of object detection tasks in aerial images, a hybrid model combining the GRoIE and Side-Aware Boundary Localization modules is presented so as to increase object detection performance in aerial images. After that, the hybrid model is trained via a Dynamic strategy (Dynamic R-CNN).

Instead of using the RoI Pooling operation to extract regions of interest from only one feature layer of the Feature Pyramid Network, the GRoIE module was employed to take advantages of RoI features from all FPN layers, leading to better performance. With extracted RoI features, the SABL mechanism handles the regression task, which predicts the exact coordinates where the objects localize on an image. SABL was designed to regress the bounding boxes' offsets via boundary features (left, right, top, down) of RoI features, which can be effective for objects which various sizes in aerial images.

Finally, the dynamic strategy was applied to effectively train the Region Proposal Network by adjusting the IoU threshold to select the highest-quality background and foreground samples. This adjustment led to a better RPN network, proposing more effective regions of interest.

By default, ResNet-101 was used as the backbone network to extract features of input images. However, the intensive experiments wereconducted on two backbones: ResNeXt-101 and ResNeSt-101, to analyze and find the best suitable for the proposed hybrid model to detect objects in aerial images. Training the Region Proposal Network used Cross-Entropy and Focal loss functions, the effectiveness of which was evaluated on the hybrid model. The pipeline of the hybrid model is demonstrated in Figure 4.

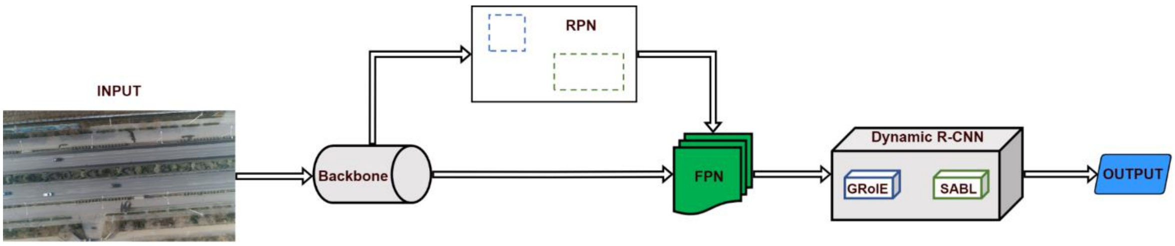


Figure 4. The pipeline of GRoIE combined with Dynamic RCNN and SABL

4. EXPERIMENT RESULTS

4.1. XDUAV dataset

XDUAV (Xie et al., 2018) was chosen as the benchmark for experiments. This dataset contains 4344 images include of 3485 training images, 869 testing images and 869 validating images, which were captured by quadcopter DJI Phantom 2 in a part of the city and countryside areas of Xi'an, China. The XDUAV dataset contains six classes, namely: "car", "bus", "truck", "tanker", "motor" and "bicycle". The class "car" has the most occurrence. The number of frequencies of each object class in the dataset is shown in Table 1.

Images have a resolution of 1920x1080. In the XDUAV dataset, the high number of small vehicles brings some of the most crucial problems such as occlusion, truncation and multiple changes in object orientation angle. Some example illustrations of the XDUAV dataset are shown in Figure 5.

4.2. Implementation Detail

The experiment was implemented on Google Colab Pro's environment. It performs on Nvidia's Tesla K8 GPU with 12GB of GDDR5 VRAM, Intel Xeon Processor with 2 cores @2.20GHz and 13 GB RAM, and the default configuration is provided by MMDetection framework 2.10.0. Experiments were mainly conducted on three backbone architectures: ResNet-101, ResNeXt-101 and ResNeSt-101 for training the GRoIE combined with Dynamic R-CNN and SABL in 24 epochs. Other methods with the same backbones were also trained on the same dataset for comparison with our approach.

4.3. Evaluation

The main metrics for evaluation of the approaches on the XDUAV dataset are AP scores.

Specifically, AP score calculation is performed on many different levels of IoU threshold which range from 50% to 95% with intermittent steps of 5%. Furthermore, AP_{50} and AP_{75} are calculated corresponding with the high thresholds of 50% and 75%.

Table 1. The frequency of each object class within the XDUAV dataset

	Category	Car	Bus	Truck	Motor	Bicycle	Tanker
XDUAV Dataset	Train	20108	1681	1608	4005	1193	110
	Test	6956	573	550	1378	426	35
	Val	6777	594	532	1273	405	28
	Total	33841	2690	2848	6656	2024	173

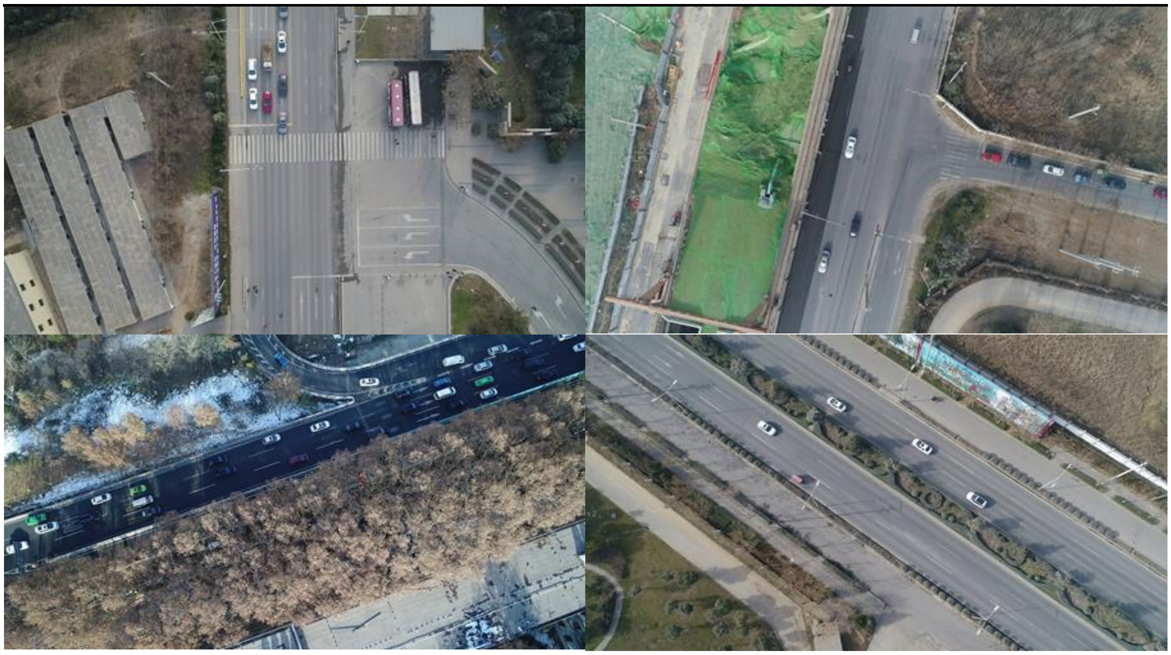


Figure 5. Sample images from the XDUAV dataset

4.4. Results Analysis

Firstly, the demonstration of the GRoIE method was divided into two versions: (1) using Cross-Entropy algorithm (CE) as Cross-Entropy loss in bounding box function on ResNet-50 and (2) using Focal-Loss (FL). The purpose was to compare the compatibilities of two loss functions on the XDUAV dataset. As the experimental result shows in Table 2, using Cross-Entropy brought higher scores in all classes of the dataset, as well as higher mAP scores than Focal-Loss. Specifically, the result of mAP for CE was 70%, which was 1.2 times higher than using FL. Moreover, the AP scores of all object classes in the dataset also improved by 1% when using CE compared to using FL. Intuitively, it is shown that Focal-Loss was not suitable for two-stage models, which is correct in this case due to the fact that our approach is an integration of Faster R-CNN. Figure 6, shows the comparison between the Cross-Entropy loss and the Focal-Loss.

In the second experiment, comparing our experimental approach and Faster R-CNN with GRoIE as RoI extractor on the same backbone, ResNet-101. Table 3 highlights that our approach achieved higher mAP and mAP75 by 1% compared with the default configuration of Faster R-CNN Our approached increased the mAP scores significantly from 82% to 86% for class “bus” and from 52% to 55% for class “motor”.

However, the mAP scores of class “tanker” and “bicycle” decreased slightly by 1%. This was due to the adjustment of the IoU threshold in Dynamic R-CNN, with the high IoU threshold, the number of regions of interest was minimized, which let the minority of class such as “tanker” or ”bicycle” showing decreased detection. Generally, our approach showed an improved result in which the mAP scores reached 71% compared to the 70% obtained from default faster R-CNN with GRoIE. The visualization in figure 7 shows the increase of our approach’s result compared to “Faster R-CNN + GRoIE” method.

Finally, as shown in Table 4, three different backbones including ResNet-101, ResNeXt-101, ResNeSt-101 were integrated into our approach. Overall, 4, the highest mAP score belonged to the backbone ResNeSt-101, at 72%. In addition, the mAP50 and mAP75 scores of ResNeSt-101 experiment have similar results to others at 93% and 84% respectively, except for the mAP75 scores in ResNet-101 at just only 83%. However, the AP scores of class “tanker” in ResNeSt-101 task were lowered by 1% than ResNeXt-101 experiment. This can be explained by the face that the ResNeSt architecture does not effectively support classes which have a lack of data. However, with backbone ResNeSt-101, other classes were slightly improved. The experiment is visualized in the form of Figure 8.

Table 2. The result of GRoIE method in two styles of Cross-Entropy loss: using Cross-Entropy algorithm (CE) or Focal Loss (FL)

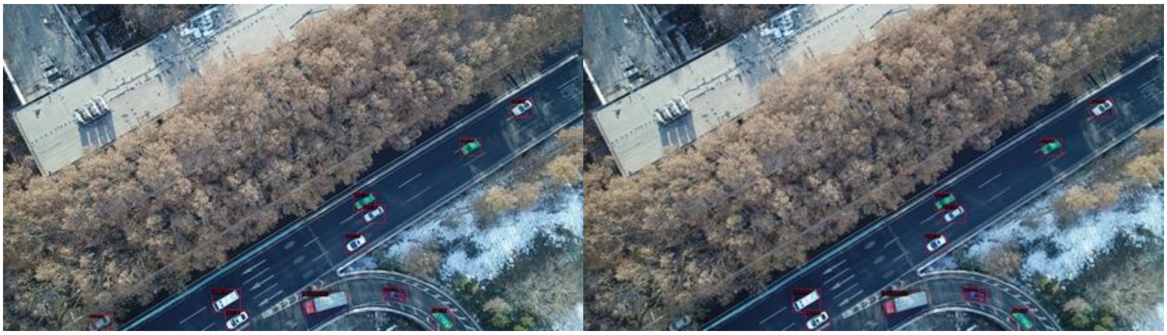
Loss function	**AP**						**mAP**	**mAP$_{50}$**	**mAP$_{75}$**
	car	**truck**	**bus**	**motor**	**bicycle**	**tanker**			
FL	0.79	0.80	0.80	0.49	0.37	0.23	0.58	0.79	0.67
CE	**0.80**	**0.81**	**0.82**	**0.55**	**0.46**	**0.46**	**0.70**	**0.93**	**0.82**

Table 3. The result between our approach and Faster RCNN+GRoIE at the same of epoch and backbone

Method	**Loss function**	**AP**						**mAP**	**mAP$_{50}$**	**mAP$_{75}$**
		car	**truck**	**bus**	**motor**	**bicycle**	**tanker**			
FasterRCNN+GRoIE	ResNet-101	0.80	0.82	0.82	0.52	**0.48**	**0.76**	0.70	0.93	0.82
Our approach	ResNet-101	**0.82**	**0.83**	**0.86**	**0.55**	0.47	0.75	**0.71**	**0.93**	**0.83**

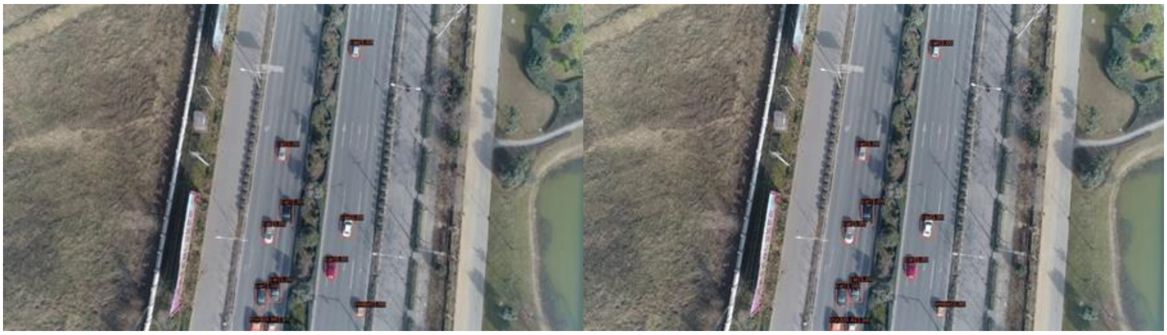
Table 4. The result of three backbone experiment on our approach at the same of epoch.

Loss function	**AP**						**mAP**	**mAP$_{50}$**	**mAP$_{75}$**
	car	**truck**	**bus**	**motor**	**bicycle**	**tanker**			
ResNet-101	0.82	**0.83**	**0.85**	0.55	0.47	0.74	0.71	**0.93**	0.83
ResNeXt-101	0.82	**0.83**	**0.85**	0.54	**0.50**	**0.76**	**0.72**	**0.93**	**0.84**
ResNeSt-101	**0.83**	**0.83**	**0.85**	**0.56**	**0.50**	0.75	**0.72**	**0.93**	**0.84**



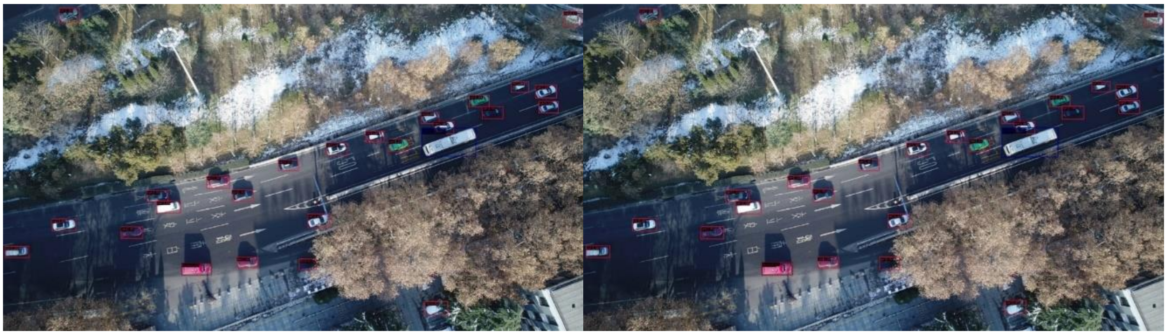
Cross-entropy experiment Focal Loss experiment

Figure 6. Visualization of the differences between Cross-Entropy experiment and Focal Loss experiment. On the left corner of Cross-Entropy experiment, the car was defined exactly, while it was undetected in the Focal Loss experiment

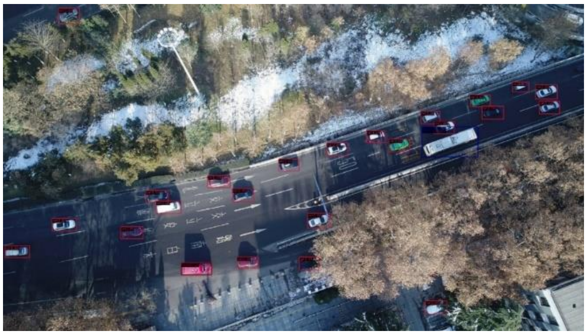


Our approach *Faster R-CNN + GRoIE*

Figure 7. Visualization of the differences between our approach and faster R-CNN with GRoIE. The orange truck at the below of image had an AP score at 0.94 in our approach, while it was only 0.42 in Faster R-CNN with GRoIE



ResNeSt-101 *ResNeXt-101*



ResNet-101

Figure 8. Visualization of the differences between three backbones: ResNet-101, ResNeXt-101 and ResNeSt-101

5. CONCLUSION

In this experiment, a significant improvement in the combination is anticipated through the implemented approach. Following rigorous testing and evaluation, it was discovered that with ResNeSt-101 as the foundation, the mAP score reached 72%, and the mAP50 and mAP75 improved as much as ResNeXt-101. Nevertheless, in addition to improving AP results in other classes, tanker AP scores decreased slightly from 76% to 75% when compared to ResNeXt-101. The quality of detection findings will be improved in the future.

ACKNOWLEDGMENT

The research team would like to express our sincere thanks to Multimedia Communications Laboratory (MMLab), University of Information Technology, Vietnam National University Ho Chi Minh City, for supporting the research.

REFERENCES

Cai, Z., & Vasconcelos, N. (2018). Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 6154-6162).

Farahnak-Ghazani, F., & Baghshah, M. S. (2016, May). Multi-label classification with feature-aware implicit encoding and generalized cross-entropy loss. In *2016 24th Iranian conference on electrical engineering (ICEE)* (pp. 1574-1579). IEEE.

Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 580-587).

Girshick, R. (2015). Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision* (pp. 1440-1448).

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, *37*(9), 1904-1916.

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).

He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision* (pp. 2961-2969).

Lin, T. Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision* (pp. 2980-2988).

Liu, S., Qi, L., Qin, H., Shi, J., & Jia, J. (2018). Path aggregation network for instance segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 8759-8768).

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems, 28. https://proceedings.neurips.cc/paper/2015

Rossi, L., Karimi, A., & Prati, A. (2021, January). A novel region of interest extraction layer for instance segmentation. In *2020 25th International Conference on Pattern Recognition (ICPR)* (pp. 2203-2209). IEEE.

Wang, J., Zhang, W., Cao, Y., Chen, K., Pang, J., Gong, T., ... & Lin, D. (2020, August). Side-aware boundary localization for more precise object detection. In *European Conference on Computer Vision* (pp. 403-419). Springer, Cham.

Wan, J., Zhang, B., Zhao, Y., Du, Y., & Tong, Z. (2021). VistrongerDet: Stronger Visual Information for Object Detection in VisDrone Images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 2820-2829).

Xie, S., Girshick, R., Dollár, P., Tu, Z., & He, K. (2017). Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1492-1500).

Xie, X., Yang, W., Cao, G., Yang, J., & Shi, G. (2018, September). *The Collected XDUAV Dataset*. Available online: https://share.weiyun.com/8rAu3kqr.

Zhang, H., Chang, H., Ma, B., Wang, N., & Chen, X. (2020, August). Dynamic R-CNN: Towards high quality object detection via dynamic training. In *European conference on computer vision* (pp. 260-275). Springer, Cham.

Zhang, H., Wu, C., Zhang, Z., Zhu, Y., Lin, H., Zhang, Z., ... & Smola, A. (2020). Resnest: Split-attention networks. *arXiv preprint arXiv:2004.08955*.