IEIE Transactions on Smart Processing and Computing

Light-Weight Deep Neural Network for Small Vehicle Detection Using Model-Scaled YOLOv4

Mingi Kim[1], Heegwang Kim[2], and *Joonki Paik[1, 2]

[1] Graduate School of Artificial Intelligence, Chung-Ang University / Seoul 06974, South Korea
[2] Department of Image, Graduate School of Advanced Imaging Science, Multimedia and Film, Chung-Ang University / Seoul 06974, Korea.

mgkim@ipis.cau.ac.kr, heegwang@ipis.cau.ac.kr, paikj@cau.ac.kr

***** Corresponding Author: Joonki Paik, paikj@cau.ac.kr

* *Regular Paper.*

***Abstract*:** In this paper, we present a light-weight deep neural network based on an efficiently scaled YOLOv4 model to detect small objects in drone images. Since a drone image has only small objects, we modified the YOLOv4 model by eliminating the Head Layer that detects large objects. Such elimination realizes a light-weight model, and reduces the processing time for Non-Maximum Suppression (NMS). In addition, the appropriately scaled model for small object detection can be mounted in a drone. In pursuit of light-weight detection of small objects with minimal performance degradation, Attention Stacked Hourglass Network (ASHN) was used for feature fusion. As a result, we proposed an efficient light-weight deep neural network model for drone environments with high small object ratio. Extensive experiments compared with the original model demonstrated that the proposed network can be extended to detect various small objects.

***Keywords*:** Small object detection, Deep learning, Light-Weight, Attention mechanism

1. Introduction

Deep learning-based object detection has been used in many applications. Among various object detection technologies, detection using a drone camera becomes popular in various fields such as military, industry, security, and transportation. In particular, drone-based traffic analysis gains increasing attention in various applications such as traffic jam identification, illegal parking detection, and intelligent traffic control systems. However, drone images are different from CCTV or vehicle camera images in that a drone image taken from high altitudes and various angles has only small objects that have different angle and scale of features. Since general object detection model requires many parameters, high power consumption, and large memory size, it is not suitable to be mounted as a low-power embedded system in a drone. To create an efficient model that can be mounted on a drone or low-power embedded system, model scaling technique can be used. To reduce the size of a model, a general model scaling technique designs changes the depth, width, and input resolution of the backbone network. Therefore, we propose a novel efficient light-weight deep neural network model for small vehicle detection in drone images. The YOLOv4-s model, which is the lightest version for real-time object detection [1], was used as the baseline model. Because relatively large objects do not appear due to the characteristics of drone images, the head layer that detects large objects has been removed. Also, light-weight was performed through efficient model scaling for small object detection. To compensate for lost information, the Attention Stacked Hourglass Network (ASHN) is added to the middle level of the backbone network where feature fusion is performed. More specifically, the ASHN was designed as a structure to supplement for features in the multi-scale and filter that were lost through the previously proposed method. The purpose of this paper is to design a model according to the limited environment or the task to be used. Through the proposed methods, we realized an efficient light-weight deep neural network model that detects small vehicles in a hardware- limited environment such as drones. Section II gives the review of a related work about small object detection and model scaling technique. Section III presents the proposed efficient light-weight deep neural network model. Section IV summarizes experimental results, and Sections V concludes the paper.

2. Related Work

2.1 Object Detection in UAV images

With the advancement of technology, unmanned aerial vehicles (UAVs) equipped with cameras can flexibly acquire ground images without geographical restrictions. Therefore, UAVs images have wide applications in detecting human, vehicles and military targets for search and rescue operations. For robust and accurate detection from UAV images in real-life environment, state-of-the-art deep learning-based object detectors should be scaled down to reduce the weight and to save the power of the UAV. We briefly discuss about cons and pros of existing detectors in the sense of the number of stages.

1) Two-Stage Detector

Regions with convolutional neural networks features (R- CNN) is the first two-stage detector that sequentially per- forms region proposal and classification. To improve the performance of R-CNN, various enhanced methods have been proposed. Faster R-CNN [2] calculates region of interest (ROI) through region proposal network (RPN) instead of using selective search. The RPN improves the accuracy of learning with accelerated ROI calculation using GPU. Cascade R-CNN [3] uses multiple classifiers, where each classifier receives a bounding box step by step and performs a new classification task under assumption that the bounding box produced at each step will be more accurate. The classifier of the next stage has a higher intersection of union (IoU) value than the previous stage. Libra R-CNN [4] refers to imbalance object detection at three levels, such as sample, feature, and objective levels, for balanced learning. The Libra R-CNN solves the imbalance problem by integrating three new components: IoU balanced sampling, balanced feature pyramid, and balanced ℓ1 loss. CRAFT [5] is a model to solve the scene text detection (STD) problem by predicting the region score of the probability that the pixel is the center of the character and the affinity score of the probability that the pixel is the center of two adjacent characters.

2) One-Stage Detector

Single shot detector (SSD) is the first one-stage detector that performs classification and regional proposal at the same time [6]. It improves the detection speed by replacing the last fully-connected (FC) layer of the network with a convolution layer. This model predicts six different scale feature maps obtained through the convolution layer in the middle of the convolutional network. SSD estimates an object using a default box with different scales and aspect ratios for each grid cell of the feature map. M2Det proposed a multi-Level feature pyramid network (MLFPN) that consists of three modules to find objects with different sizes and complexity of appearance [7]. The feature fusion module (FFM) creates an optimal feature by fusing shallow and deep features from the backbone. Thinned U-shape module (TUM) is reconstructed using the second version of the feature fusion module (FFMv2). M2Det uses multi-level, multi-scale features through scale-wise feature concatenation and channel-wise attention. This model has an end-to-end form combining MLFPN and SSD. YOLOv4 combines CSPDarknet53-based backbone architecture with spatial pyramid pooling (SPP) [8] and path aggregation network (PAN) [9] to enable fast learning and inference as well as high performance with a single GPU [1]. In addition, various data augmentation techniques were presented to improve the detection performance without increasing inference time.

In order to be mounted on a real drone and used in a low-power embedded system, a fast, light-weight detector is needed. However, widely used one-stage detector models were generally tested by MSCOCO [10], and it is difficult to applied to UAV images since object detection in UAV images taken at high altitude and at various angles has low accuracy due to complex background, small size of the object, and object change according to angle. Therefore, small object detection suitable for UAV images is an open, challenging issue.

2.2 Small Object Detection

Detection of small objects is a difficult task in the field of computer vision due to a very small number of pixels in the object and imbalanced amount of information between background and objects. Among various approaches to detect small objects, it is a common practice to make the CNN layer deeper to obtain the higher-level feature map containing semantic information of the object at the cost of losing a low-level spatial information. To solve this problem, various researches combining shallow and deep features have been proposed [6], [9], [11]–[14]. The combined method can learn shallow level features even in the deeper layer. A limited amount of contextual information in a small object, which is as small as 32 32 pixels, is another challenge [15] since local context contains very important information such as edge, color, and texture of an object. To compensate for the local pixel context, the filter size of the network was increased or a deconvolution layer was added for a higher- level feature map of the image [16]–[18].

Recently, multi-scale feature maps are widely used. How- ever, the matching ratio between the feature map and the ground truth small object is still insufficient since the anchor was not appropriately adjusted. The low matching ratio makes the performance of small object detection lower than that of large objects. In order to solve the imbalance of small objects, methods for generating positive examples for small objects using multi-scale feature map and anchor box have been proposed [14], [19]–[21]. Only anchors with high IoU scores are designated as positive examples, and all others are considered to be negative, which results in a severe imbalance between positive and negative examples. To

solve this problem, weights of the network adjusted so that the positive and negative examples have similar numbers after training the machine learning model based on the data distribution [4], [6], [22]–[25]. Another approach designs a new loss function to reset the weights between positive and negative example data that are unbalanced for each training [26], [27].

2.3 Model Scaling

Model Scaling refers to a technique that determines the size of a model by changing the width, depth, and resolution, which are factors that determine the size and amount of calculation of a baseline model [28], [29]. In Scaled- YOLOv4 [30] and EfficientDet [31], model scale-up was applied using these techniques. Width scaling changes the number of filters (channels). It is a common understanding that the wider network can extract the finer information. On the other-hand depth scaling changes the number of layers. Finally, resolution scaling changes the resolution of the input image. In EfficientDet, various conditions for model scaling were tested. Increased width or depth makes convergence earlier whereas increased resolution makes accuracy higher. In other words, in model scaling, the change in resolution has a great effect on the performance. In particular, it was proved that raising three elements at the same time has the best performance. Therefore, Scaled-YOLOv4 and EfficientDet used model scaling technique to fix base model and adjusted three elements to fit the model size through the factor value. On the contrary, through scale-down of the model scaling technology, it is possible to create a light-weight model that is smaller than the base model. It uses three elements like scale- up, but the model scale-down method that can minimize performance degradation is also a light-weight method.

3. The Proposed Method

Although state-of-the-art object detection models show significantly improved performance, it is difficult to embed these models in an actual drone due to the limited hardware capacity. Therefore, a light-weight network with an acceptable performance is essential for a low-power embedded system such as a drone. In this paper, we present an efficiently scaled YOLOv4-s model [1] to detect small objects in drone images without significantly losing the detection performance.

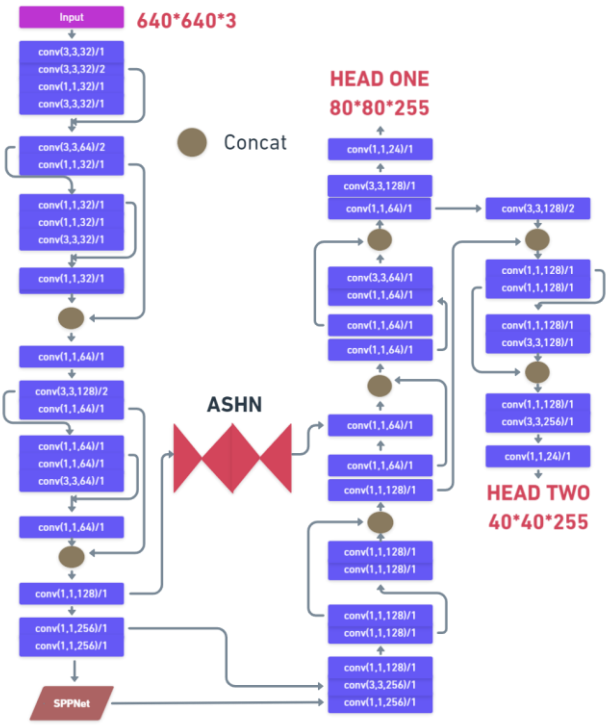

Figure 1. Architecture of the proposed model

3.1 Removing the Head Layer for Small Object

The baseline YOLOv4-s has three head layers [1]. In a 640x640 input image, small-sized objects are detected in 80x80 feature maps, medium-sized objects in 40x40 feature maps, and large objects in 20x20 feature maps. However, since drone images are taken at high altitudes, most objects fall in categories of small or medium-sized objects. Since most objects in a drone image fall in categories of small and medium objects, we removed the 20x20 feature map head layer for large objects together with the connected neck layer. As a result, the size of weight and processing time are reduced for the NMS process. The reduced model has six anchor boxes in two head layers whereas the original model has nine anchor boxes in three head layers.

3.2 Model Scaling

Figure 1 shows the architecture of the proposed model, where we reduced the size of weight by scaling-down the baseline model. The resolution of an input image is fixed to 640 × 640 without scale-down, whereas the model is scaled down along both width and depth. While scaling down the width and depth, we tested the performance change according to each element scaling. As the training data, we used Visdrone2019-Det [32] that obtained from an actual drone. Visdrone2019-Det consists of 10 classes under various environments. Since the objective of this work is to

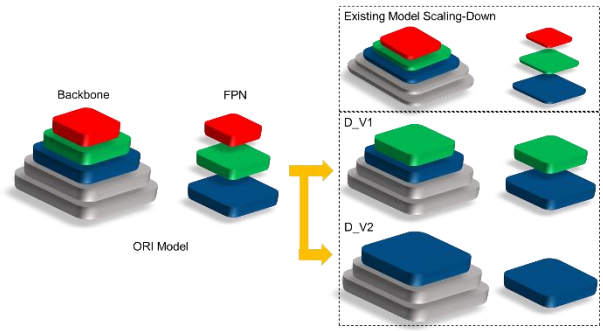

Figure 2. Depth Scaling. Maintaining the structure of the model: Existing Model Scaling Down and change the structure of the model: D_V1, D_V2 Model. The thickness of each feature map means the number of layers of the feature map.

detect small vehicles, we conducted an experiment to find the reference points of depth scaling and width scaling using only Car, Bus, and Truck (Vehicle) classes among Visdrone2019-Det. The proposed scaling method is different from existing methods in that we change the model structure by estimating the scaling levels of depth and width that are effective for small object feature learning.

1) Model Depth Scaling

Figure 2 shows the depth scaling of the proposed network. In order to find the reference point of depth scaling, we compared the D-V1 model in which the layer was removed as the starting point in the 20x20 feature map and the D-V2 model as the starting point in the 40x40 feature map with the original model. Each model adds 1 or 2 convolution layers to match the feature map size when performing feature fusion between the neck and head layers. Table 1 shows the result about the depth scaling. Compared with the original model, the depth scaling model recorded higher performance with reduced parameters since most of detection networks have global and spatial information at the shallow level while specific and semantic information as the size of the feature map is reduced at the deep level. However, the feature information of small objects that can be learned is very small. Therefore, if the layer becomes deeper, the more down- sampling is required in the feature map, and small object feature information may be lost. It can lead to confusion in the training process. In that case, efficient learning is not possible, and unnecessary layers and filters remains with high computational load. As can be seen from this experiment, we observed that if the deep level layer with the reduced size of the feature map was removed, the features of small objects could be efficiently learned and the recall increased. We also set a standard for efficiently learning the feature information of small objects. Therefore, the network proposed in this paper selects the depth scaling point as D-V2.

2) Model Width Scaling

For width scaling, we compared the W-V1 model with the maximum filter limit of 256 and the W-V2 model with the maximum filter limit of 128 with the original model. Table 1 shows the result of width scaling. In the case of width scaling, the number of model filters has a large effect on the model size. And the lighter the model, the more model filter affects the model size. It is clear that the more the number of filters, the more detailed information can be learned. However, it is important to understand the minimum number of optimal filters when the light-weight is required. As a result of the width scaling experiment, we observed that the more scale-down, the lower the performance. Compared with the original model, the W-V1 model becomes slightly lighter while preserving a similar performance. However, although the performance of W-V2 decreased slightly, it shows a significant light-weight effect. In the case of width scaling, the W-V1 model should be adopted in terms of performance, but the W-V2 model can show good value in terms of weight reduction. Therefore, we combined W-V1 and W-V2 with depth scaling, and tested compound scaling. Both W-V1 and W-V2 models were combined with depth scaling to experiment with compound scaling.

Table 1. Model Scaling Results

Model	Parameters	Precision	Recall	mAP@.5	mAP@.5:.95
YOLOv4-s(ORI)	8.06MB	0.426	0.603	0.558	0.367
Depth Scaling					
D-V1	6.28MB	0.402	0.595	0.556	0.368
D-V2	6.40MB	0.414	0.62	0.57	0.380
Width Scaling					
W-V1	6.48MB	0.392	0.609	0.555	0.365
W-V2	2.73MB	0.396	0.594	0.549	0.358
Compound Scaling					
D-V2+W-V1	5.80MB	0.417	0.628	**0.579**	**0.39**
D-V2+W-V2	2.42MB	0.386	0.612	0.555	0.369
Compound Scaling + Head Layer Elimination					
D-V2+W-V1 +HLE	2.55MB	0.490	0.580	**0.568**	**0.387**
D-V2+W-V2 +HLE	**1.59MB**	0.468	0.583	0.556	0.374

3) Model Compound Scaling

Both W-V1 and W-V2 models of width scaling were combined with the D-V2 model adopted in depth scaling for compound scaling. The experimental result of compound scaling is shown in Table 1. Although the weight was reduced by scaling down depth and width, the overall performance was maintained or even improved because of an efficient model structure for small object detection by eliminating unnecessary layers and filters. The result of removing the Head Layer in the first method to the above two models is combined scaling and head layer elimination in Table 1. By applying model scale-down and head layer elimination method, significant light-weight was achieved. Also, an important point to consider through the two experimental results is that performance was improved only through model scaling. This proves that the structure of the model is important in small object detection task.

4) Model Scaling Network

Based on the experimental results, the proposed network removes the deep convolution layers at the bottom of the network from the starting point (D-V2) where the feature map becomes 40x40 in the baseline network. In addition, the network was scaled-down by combining the width scaling model (W-V1) that limits the number of filters in the network layer up to 256. However, in terms of weight reduction, it can be a good option to using the model combination of D-V1 and W-V1. The result of a compromise between light-weight and performance is the former model. In conclusion, we designed a network that maximizes object location information and global features by minimizing the reduction of feature maps through depth scaling. Also, it is designed to efficiently use the minimum number of filters through width scaling. The proposed deep neural network applies model scaling method to the depth and width of the backbone network to construct an efficient and low-power-embedded network suitable for the small object detection task. However, various feature information was lost because of model scaling.

3.3 Attention Stacked Hourglass Network

Attention Stacked Hourglass Network (ASHN) was added to compensate for feature information loss and to effectively detect small objects. ASHN is shown in Figure 3. The existing Hourglass Network [33] is used to Human Pose Estimation and can extract various feature information through iterative dimensionality reduction and dimensionality increase on multiple scales. In addition, by combining feature maps, it is possible to re-estimate the features of the overall image. Stacked Hourglass Network [33] is a structure in which multiple Hourglass Networks are stacked. The structure of Hourglass Network is used to compensate for the feature information loss of depth and width scaling during model scaling. In order to focus on the weight of the small objects features, the attention mechanism is added to the Hourglass Network to design the Attention Hourglass Network as shown in Figure 3.

1) Feature Fusion using Attention Stacked Hourglass Network

In the backbone network, each Shallow, Medium, and Deep level has different feature information. M2Det [7] shows that traffic-sign, car, and pedestrian objects have different feature characteristics and are detected at different levels and scales. Since the object in the drone image is almost small size and contains little object feature information, it would be better to focus on smaller feature information. In general, we know that low-level features are useful for detecting small objects. In order to detect small objects, not only a low-level feature but also a high-level feature that captures the context of the image are required. Therefore, if low-level and high-level feature information are used, the detection performance will be better. Therefore, Feature Fusion was conducted by using Attention Stacked Hourglass Network to extract small object features from the network.

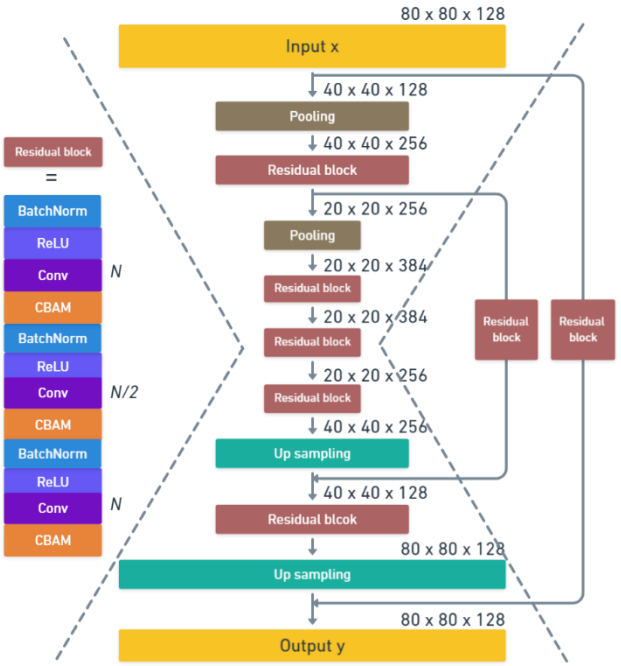

Figure 3. Attention Stacked Hourglass Network

2) Attention Stacked Hourglass Network Structure

Two things were considered when designing the attention stacked hourglass network as shown in Figure 2. First, it was designed to supplement the feature information loss that can be learned from various feature map sizes through depth scaling. Second, it is designed to supplement filter information that can learn various and fine features through width scaling. ASHN is designed as follows. Hourglass Network was stacked 2 times, and features were extracted from 80x80, 40x40, and 20x20 feature map sizes through max pooling twice. Also, the filter is designed to increase by 128 whenever the feature map size decreases. What we aimed is that extract features specified at various scales with shallow level feature information. It is designed to compensate for feature loss in various feature map sizes through backbone depth scaling with repeated dimensionality reduction and increase through hourglass network. By increasing the number of filters in the hourglass network, the various information of missing filters through backbone width scaling could be supplemented by feature fusion with richer information. Additionally, in order to effectively learn small object features, ASHN was designed to focus on the weights of small objects by adding an attention module of CBAM [34] structure in the residual block. Attention module is largely divided into Channel Attention Module [35] and Spatial Attention Module [36]. Channel attention extracts only the critical part of the feature from each channel and extracts one value for each channel. By taking the corresponding value as a sigmoid and producing a product on the input feature map, the values to be paid attention to in each channel have a high weight. Spatial attention creates a feature map with one channel

through max and average pooling of all channels. If the corresponding feature map is taken as a sigmoid and multiplied by the input feature map, the values to be paid attention to in the entire channel have a high weight. In conclusion, channel attention pays attention to 'what', and spatial attention pays attention to 'where'. CBAM [34] applied to our network is a module that sequentially processes Channel Attention and Spatial Attention. Table 2 shows the result of adding ASHN to the two models adopted in Method 2. In Table 2, it can be con- firmed that although the weight has been reduced compared to the original model, there is an improvement in performance. Through feature fusion using ASHN, it contributed to performance improvement by concentrating weights on features of small objects. In particular, it can be seen that the precision has significantly increased. Small objects can be confused with the background and other objects because the size of the object is so small. However, the proposed network in this paper tried to solve this problem and the precision was greatly improved.

Table 2. Compound + Head Layer Elimination + ASHN Experimental Results

Model	Parameters	Precision	Recall	mAP@.5	mAP@.5:.95
YOLOv4-s(ORI)	8.06MB	0.426	0.603	0.558	0.367
D-V2+W-V1 +HLE+ASHN	**5.73MB**	**0.534**	**0.60**	**0.591**	**0.399**
D_V2+W_V2 +HLE+ASHN	4.77MB	0.502	0.579	0.571	0.384

Table 3. Various Dataset Experimental Results

Model	Parameters	Precision	Recall	mAP@.5	mAP@.5:.95
Visdrone-DET2019 [32]					
YOLOv4-s(ORI)	8.06MB	0.426	0.603	0.558	0.367
Our Network	**5.73MB**	**0.534**	**0.60**	**0.591**	**0.399**
UAVDT [37]					
YOLOv4-s(ORI)	8.06MB	0.655	0.991	0.992	0.747
Our Network	**5.73MB**	**0.785**	**0.991**	**0.992**	**0.774**
CARPK [38]					
YOLOv4-s(ORI)	8.06MB	0.505	0.996	0.996	0.819
Our Network	**5.73MB**	**0.682**	**0.997**	**0.997**	**0.844**

Table 4. Visdrone-DET2019 10 Class Experimental Results

Model	Parameters	Precision	Recall	mAP@.5	mAP@.5:.95
YOLOv4-s(ORI)	8.08MB	0.296	0.426	0.364	0.213
Our Network	**5.74MB**	**0.372**	**0.429**	**0.391**	**0.233**

4. Experimental Result

The experiment was implemented in the RTX 3090 (24G) environment. YOLOv4-s (Ori) and Our Network model were tested under the same environmental conditions with 16 batch size, learning_rate 0.00261, and 640x640 image size. Through k-means, the original model was trained by setting 9 anchor boxes and Our Network model by setting 6 anchor boxes. The experimental results of the previous experiment were trained and evaluated using Vehicle (Car, Bus, Truck) among 10 classes of Visdrone2019-Det [32] for the purpose of Small Vehicle Detection. Additionally, to validate the performance of Our Network in various environments, two additional drone environment public data sets (UAVDT[37], CARPK [38]) were additionally tested. Visdrone-DET2019 and UAVDT used 3 classes of Car, Bus, and Truck, and CARPK used 1 class of Car. Table 3 shows the comparison results of Parameters, Precision, Recall, and mAP for each data set. Our Network has 5.73 (MB) parameters, which is about 1.4 times lighter than the original YOLOv4-s (ORI) model. In addition, high mAP was recorded by supplementing small object feature information in the image through Feature Fusion using Attention Stacked Hourglass Network. In particular, we can see that Precision has a lot of difference compared to other performance indicators. Precision is an indicator of false detection. Our Network showed more robust results against false detection than the original model. In addition, Our Net- work is not limited to small vehicles, but is also sufficiently applicable to small objects of various classes. Table 4 shows the experimental results for 10 classes (Pedestrian, Person, Car, Van, Bus, Truck, Motor, Bicycle, Awning-tricycle, and Tricycle) used in the Visdrone-DET2019 challenge. The results obtained through this experiment are as follows. When designing a model for a small object detection task, the deep depth of the model may confuse the learning of features of small objects. Therefore, it can be seen that the scale of depth plays an important role in Small Object Detection Task. In addition, it can be seen that feature fusion using Attention Stacked Hourglass Network contributes to the improvement of model performance by extracting small object features of various scales and various filters. The accuracy and model size of the general object detection model have a proportional relationship. However, it is impossible to build a model with good performance in all tasks. The contribution in this paper is to suggest a method to efficiently construct a model that can detect small objects and mount them in a low-embedded environment. A more efficient model can be designed by constructing a model differently for each domain, application, and class. Figure 4 is a comparison result image. Figure 4 is the experimental result of Visdrone-DET2019 Test. Car is expressed in Red, Truck in Green, and Bus in Blue. It performed better than the original model at various angles, altitudes, and illuminance. As shown in Table 3, Our Network is robust against misclassification. In the low-light environment, the original model has misclassification and non-detection for classification, but Our Network correctly detected it. Also, unlike Ori Model, which detects the back- ground as a class at high altitude, Our Network detects small objects well.

Figure 4. Experiment Results. The following is the experimental result of Visdrone-DET2019 test. Each image consists of 3 pairs. The top is Ground Truth (GT), the middle is the Original Model, and the bottom is Our Network.

5. Conclusion

In this paper, we proposed an efficient light-weight and deep neural network model for small vehicle detection in a drone environment. Considering the drone environment with a high proportion of small objects, light-weight was performed by applying Head layer Elimination to detect large size objects and efficient model scaling. In addition, we supplemented the lost information with Feature Fusion using Attention Stacked Hourglass Network and focused on feature information on small objects. As a result, the model parameters were reduced by 1.4 times compared to the original model. Also, mAP recorded higher performance. The direction pursued by this paper is to efficiently detect small objects and efficiently build a model to mount them in a limited environment (low- power embedded). We constructed a model for detecting small vehicles in a drone environment with limited computational amount. The model can be applied to various places such as Identification traffic jam, illegal parking detection, and intelligent traffic system. In addition, it can be applied to various systems such as CCTV and portable cameras, which are low-embedded environments as well as drones. Lastly, our network is not limited to vehicles and can be used for small object detection tasks of various classes.

6. Acknowledgement

This work was supported by the Institute of Information & communications Technology Planning & Evaluation (IITP) grant, which is funded by the Korean government (MSIT) (2021-0-01341, Artificial Intelligence Graduate School Program (Chung-Ang University)), and financially supported by the Institute of Civil-Military Technology Cooperation Program funded by the Defense Acquisition Program Administration and Ministry of Trade, Industry and Energy of Korean government under grant No. UM20311RD3.

References

[1] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, “Yolov4: Optimal speed and accuracy of object detection,” arXiv preprint arXiv:2004.10934, 2020. Article (CrossRef Link)

[2] S. Ren, K. He, R. Girshick, and J. Sun, “Faster r-cnn: Towards real- time object detection with region proposal networks,” Advances in neural information processing systems, vol. 28, pp. 91–99, 2015. Article (CrossRef Link)

[3] Z. Cai and N. Vasconcelos, “Cascade r-cnn: Delving into high quality object detection,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 6154–6162. Article (CrossRef Link)

[4] J. Pang, K. Chen, J. Shi, H. Feng, W. Ouyang, and D. Lin, “Libra r-cnn: Towards balanced learning for object detection,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 821–830. Article (CrossRef Link)

[5] Y. Baek, B. Lee, D. Han, S. Yun, and H. Lee, “Character region aware- ness for text detection,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 9365–9374. Article (CrossRef Link)

[6] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, “Ssd: Single shot multibox detector,” in European conference on computer vision. Springer, 2016, pp. 21–37. Article (CrossRef Link)

[7] Q. Zhao, T. Sheng, Y. Wang, Z. Tang, Y. Chen, L. Cai, and H. Ling, “M2det: A single-shot object detector based on multi-level feature pyramid network,” in Proceedings of the AAAI conference on artificial intelligence, vol. 33, no. 01, 2019, pp. 9259–9266. Article (CrossRef Link)

[8] P. Purkait, C. Zhao, and C. Zach, “Spp-net: Deep absolute pose regression with synthetic views,” arXiv preprint arXiv:1712.03452, 2017. Article (CrossRef Link)

[9] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, “Path aggregation network for instance segmentation,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 8759–8768. Article (CrossRef Link)

[10] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, “Microsoft coco: Common objects in context,” in European conference on computer vision. Springer, 2014, pp. 740–755. Article (CrossRef Link)

[11] Z. Cai, Q. Fan, R. S. Feris, and N. Vasconcelos, “A unified multi-scale deep convolutional neural network for fast object detection,” in European conference on computer vision. Springer, 2016, pp. 354–370. Article (CrossRef Link)

[12] C.-Y. Fu, W. Liu, A. Ranga, A. Tyagi, and A. C. Berg, “Dssd: Deconvolu- tional single shot detector,” arXiv preprint arXiv:1701.06659, 2017. Article (CrossRef Link)

[13] T. Kong, A. Yao, Y. Chen, and F. Sun, “Hypernet: Towards accurate region proposal generation and joint object detection,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 845– 853. Article (CrossRef Link)

[14] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, “Feature pyramid networks for object detection,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2117–2125. Article (CrossRef Link)

[15] Y. Liu, P. Sun, N. Wergeles, and Y. Shang, “A survey and performance evaluation of deep learning methods for small object detection,” Expert Systems with Applications, p. 114602, 2021. Article (CrossRef Link)

[16] R. Girshick, J. Donahue, T. Darrell, and J. Malik,

"Rich feature hierarchies for accurate object detection and semantic segmentation," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 580–587. Article (CrossRef Link)

[17] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," IEEE transactions on pattern analysis and machine intelligence, vol. 40, no. 4, pp. 834–848, 2017. Article (CrossRef Link)

[18] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," arXiv preprint arXiv:1511.07122, 2015. Article (CrossRef Link)

[19] Y. Li, Y. Chen, N. Wang, and Z. Zhang, "Scale-aware trident networks for object detection," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 6054–6063. Article (CrossRef Link)

[20] B. Singh and L. S. Davis, "An analysis of scale invariance in object detection snip," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 3578–3587. Article (CrossRef Link)

[21] B. Singh, M. Najibi, and L. S. Davis, "Sniper: Efficient multi-scale training," arXiv preprint arXiv:1805.09300, 2018. Article (CrossRef Link)

[22] M. Kisantal, Z. Wojna, J. Murawski, J. Naruniec, and K. Cho, "Augmentation for small object detection," arXiv preprint arXiv:1902.07296, 2019. Article (CrossRef Link)

[23] B. Zoph, E. D. Cubuk, G. Ghiasi, T.-Y. Lin, J. Shlens, and Q. V. Le, "Learning data augmentation strategies for object detection," in European Conference on Computer Vision. Springer, 2020, pp. 566–583. Article (CrossRef Link)

[24] A. Shrivastava, A. Gupta, and R. Girshick, "Training region-based object detectors with online hard example mining," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 761– 769. Article (CrossRef Link)

[25] Y. Cao, K. Chen, C. C. Loy, and D. Lin, "Prime sample attention in object detection," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 11 583–11 591. Article (CrossRef Link)

[26] K. Chen, J. Li, W. Lin, J. See, J. Wang, L. Duan, Z. Chen, C. He, and J. Zou, "Towards accurate one-stage object detection with ap-loss," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 5119–5127. Article (CrossRef Link)

[27] Q. Qian, L. Chen, H. Li, and R. Jin, "Dr loss: Improving object detection by distributional ranking," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 12 164–12 172. Article (CrossRef Link)

[28] P. Dollár, M. Singh, and R. Girshick, "Fast and accurate model scaling," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 924–932. Article (CrossRef Link)

[29] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in International Conference on Machine Learning. PMLR, 2019, pp. 6105–6114. Article (CrossRef Link)

[30] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "Scaled-yolov4: Scaling cross stage partial network," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 13 029–13 038. Article (CrossRef Link)

[31] M. Tan, R. Pang, and Q. V. Le, "Efficientdet: Scalable and efficient object detection," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 10 781–10 790. Article (CrossRef Link)

[32] D. Du, P. Zhu, L. Wen, X. Bian, H. Lin, Q. Hu, T. Peng, J. Zheng, X. Wang, Y. Zhang et al. ,"Visdrone-det2019: The vision meets drone object detection in image challenge results," in Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, 2019, pp. 0–0. Article (CrossRef Link)

[33] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in European conference on computer vision. Springer, 2016, pp. 483–499. Article (CrossRef Link)

[34] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in Proceedings of the European conference on computer vision (ECCV), 2018, pp. 3–19. Article (CrossRef Link)

[35] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in Pro- ceedings of the IEEE conference on computer vision and pattern recogni- tion, 2018, pp. 7132–7141. Article (CrossRef Link)

[36] J. Park, S. Woo, J.-Y. Lee, and I. S. Kweon, "Bam: Bottleneck attention module," arXiv preprint arXiv:1807.06514, 2018. Article (CrossRef Link)

[37] D. Du, Y. Qi, H. Yu, Y. Yang, K. Duan, G. Li, W. Zhang, Q. Huang, and Q. Tian, "The unmanned aerial vehicle benchmark: Object detection and tracking," in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 370–386. Article (CrossRef Link)

[38] M.-R. Hsieh, Y.-L. Lin, and W. H. Hsu, "Drone-based object counting by spatially regularized regional proposal network," in Proceedings of the IEEE international conference on computer vision, 2017, pp. 4145–4153. Article (CrossRef Link)

Mingi Kim was born in Okcheon, Korea, in 1996. He received a B.S. degree in data analysis from Hannam University, South Korea, in 2021. He is currently pursuing an M.S. degree with the Department of Artificial Intelligence, Chung-Ang University.

Heegwang Kim was born in Seoul, Korea, in 1992. He received a B.S. degree in electronic engineering from Soongsil University, Korea, in 2016. He received an M.S. degree in Image Science from Chung-Ang University, Korea, in 2018. Currently, he is pursuing a Ph.D. degree in image engineering at Chung-Ang University.

Joonki Paik was born in Seoul, South Korea, in 1960. He received a B.S. degree in control and instrumentation engineering from Seoul National University in 1984 and M.Sc. and Ph.D. degrees in electrical engineering and computer science from Northwestern University in 1987 and 1990, respectively. From 1990 to 1993, he joined Samsung Electronics, where he designed image stabilization chipsets for consumer camcorders. Since 1993, he has been a member of the faculty of Chung-Ang University, Seoul, Korea, where he is currently a professor with the Graduate School of Advanced Imaging Science, Multimedia, and Film. From 1999 to 2002, he was a visiting professor with the Department of Electrical and Computer Engineering, University of Tennessee, Knoxville. Since 2005, he has been the director of the National Research Laboratory in the field of image processing and intelligent systems. From 2005 to 2007, he served as the dean of the Graduate School of Advanced Imaging Science, Multimedia, and Film. From 2005 to 2007, he was the director of the Seoul Future Contents Convergence Cluster established by the Seoul Research and Business Development Program. In 2008, he was a full-time technical consultant for the System LSI Division of Samsung Electronics, where he developed various computational photographic techniques, including an extended depth of field system. He has served as a member of the Presidential Advisory Board for Scientific/Technical Policy with the Korean Government and is currently serving as a technical consultant for the Korean Supreme Prosecutor's Office for computational forensics. He was a two-time recipient of the Chester-Sall Award from the IEEE Consumer Electronics Society, the Academic Award from the Institute of Electronic Engineers of Korea, and the Best Research Professor Award from Chung-Ang University. He has served the Consumer Electronics Society of the IEEE as a member of the editorial board, vice president of international affairs, and director of sister and related societies committee.