DDConv: Dilated Depthwise Convolution with YOLOv5 for Drone Imagery*

Jehwan Choi[1], Minseung Kim[2], Donggue Kim[2], and Kanghyun Jo[1]

[1] Department of Electrical, Electronic and Computer Engineering, University of Ulsan, Ulsan, South Korea
[2] School of Electrical Engineering University of Ulsan, Ulsan, South Korea
jhchoi@islab.ulsan.ac.kr, kmsoiio@naver.com, kdk6859@gmail.com, ac ejo@islab.ulsan.ac.kr
https://islab.ulsan.ac.kr/

Abstract. Unmanned aerial vehicle (UAV) with convolutional neural network (CNN)-based artificial intelligence technologies have recently received high attention in many applications. In this paper, we focus on object detection network research for drone-based real-time systems. So, the DDConv block is proposed considering the characteristics of drones such as wide shooting range, objects of various types and scales, and high resolution. DDConv analyzes images using dilated convolution and depth-wise convolution. The proposed module is applied instead of the C3 module of the YOLOv5 backbone. As a result of experiments with the proposed network, the number of parameters and the GFLOPS value decreased by about 20%. The object detection time was recorded at 6.5 ms per image. This is almost twice as fast as the original network. Although the accuracy decreased slightly, the detection result images found most of the objects well. We will apply this network to our future work like traffic analysis and surveillance systems.

Keywords: Object detection · drone dataset · dilated convolution · depth-wise convolution.

1 Introduction

Recently, drone combined with computer vision-based artificial intelligence technology has been used in various fields such as drone autonomous flight, unmanned delivery, and missing person search. The most important part of the technologies required to increase utilization in the mentioned fields is object recognition. Because the camera mounted on the drone must be able to analyze the environment around the drone and detect and avoid obstacles in its path.

* This results was supported by "Regional Innovation Strategy (RIS)" through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(MOE)(2021RIS-003).

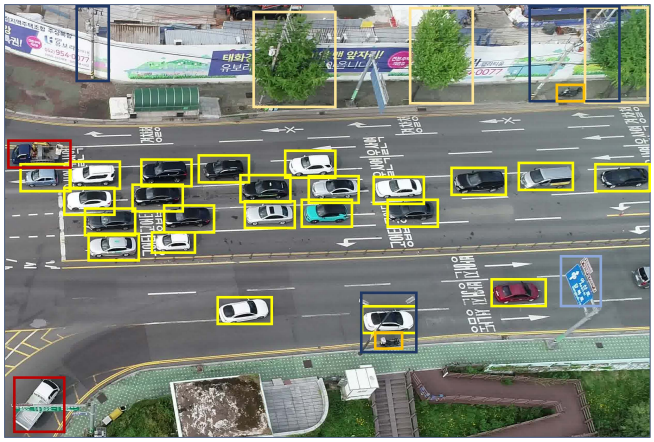
Therefore, many researchers are focusing on real-time object detection technology in drone images. CNN modification, algorithm generation, and modification methods for real-time object detection based on drone images are introduced in [1–3]. In addition, multi-object detection methods [4–7] and small object detection strategies [8–11] are also hot topics for many researchers. In this paper, we conduct a study focused on a drone image real-time object detection network which is the first step for drone traffic analysis and surveillance system.
First, we introduce the data analysis performed to create a network suitable for real-time object detection in drone images. The drone is a BEV(Bird's-Eye View) video that takes the ground from a high altitude. So, as shown in Figure 1, there are three challenges. First, the types of objects shown are very diverse because a wide area was filmed. As you can see, in the introduced Figure 1-(a), there are 6 types of objects. Also, the number of objects is very large. Second, the shape of objects is irregular even though they are in the same area because the drone is moving while filming as shown in Figure 1-(b). The third is the diversity of object forms. Similar to the second feature, but distinctly different. The forms of an object may change due to a difference in altitude, angle, and field of view like Figure 1-(c).

In order to solve mentioned problem, this paper applies a convolution technique that can effectively analyze drone images to an object detection network. First, if an image of a large area is calculated with a large filter, the probability of high-accuracy results increases but the amount of calculation increases significantly. Therefore, we apply dilated convolution technique which produces an effect similar to a large-sized filter while maintaining the amount of computation. In addition, a depth-wise convolution technique that has a similar effect to normal convolution but with a reduced amount of computation is applied to secure real-time. Because when the amount of computation is reduced, the time to detect objects is reduced. Second, the number of detection heads is increased from 3 to 4 according to the scale of the object in the drone image. As shown in Figure 2, objects in drone images can be divided into four scales. Therefore, the detection head must be configured according to the scale of objects in the drone image in order to find many objects of various scales.

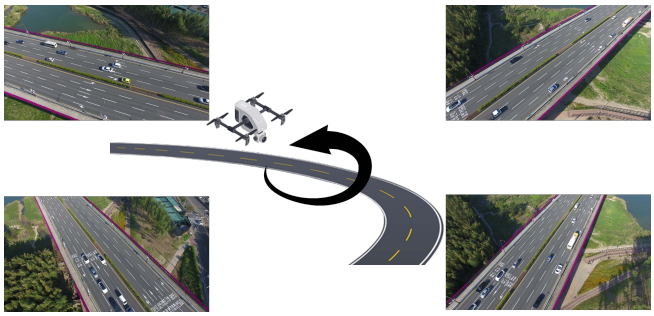
2 Related work

2.1 CNN-based object detection

Object detection is a computer technology related to computer vision and video processing to determine the existence of an object. Typically, we use a method of pre-extracting the features of an object and detecting them within a given image. In computer vision, there is a technique for object detection in a CNN method, which can maintain spatial information of an image and identify feature points using convolution. Among the CNN techniques, YOLOv5 [12] is one of the many methods used for real-time object detection what is popular among the YOLO series and has the advantage of being fast in computational speed, so it used for real-time object detection. The YOLOv5 series includes nano, small,



(a)



(b)

(c)

Fig. 1: Challenged of drone image object detection, (a) Existence of many types and number of objects in one image, (b) Difficulty in analyzing same-area information due to the irregular movement path of the drone, (c) the diversity according to the taking altitude, angle, and field of view in the same object

Fig. 2: Illustration of object scale comparison in drone images.

medium, large, and x-large. And they have difference in accuracy and computational speed by adjusting the number of convolution times in the backbone. Instead of changing the performance according to the number of convolutions, it can affect the performance through a change in the convolution method, such as Transpose convolution, Separable convolution, Dilated convolution, Depth-wise Separable convolution, etc. Dilated convolution was selected as method for efficiently learning wide-pixel photos. Dilated convolution is the main idea in the DilatedNet [13] that increases the reception file by adding zero padding inside the convolution filter, allowing the same number of input pixels while accommodating a wide range of inputs. This paper [14] also improves performance by making the CFEM(Context Feature Enhancement Module) a multi-path dilated convolutional layer.

2.2 High speed CNN method

There are two key evaluation indicators for CNN. The first is accuracy, and the second is speed. The topic of this paper is real-time object detection in drone images, and it focuses on speed because it is a priority to detect many objects quickly. The Depth-wise Separable Convolution used as a method for improving CNN speed performance. The paper MobileNets [15] presents the Depth-wise

Separable Convolution method. This paper is a representative paper that speeds up CNN by changing the calculation method without focusing on reducing the amount of calculation by increasing the filter size as 5x5 and 7x7. Therefore, Our paper applied the Depth-wise Separable Convolution method to speed up the CNN computation.

2.3 Object detection using drone image

When using drone dataset, the background occupies a large portion because it was shoted at a high altitude, and the size of objects to be detected is small. As a way to increase the accuracy of small object detection, there is a method of constructing a dataset using OBB(Object Bounding Box) in the paper [16], as shown in Fig1. In other ways, it's easier to find large objects in small images, and it's easier to find small objects in large images. For example, YOLOv1 [17] is fast but has a low ability to detect small things, and YOLOv3 [18] has increased its ability to catch small objects by performing prediction on three scales. This paper got the idea from this method and used the dilated convolution to YOLOv5's backbone to obtain a wide range of values, which would help detect small objects. In fact, it worked in a paper [19] that used dilated convolution to detect small objects.

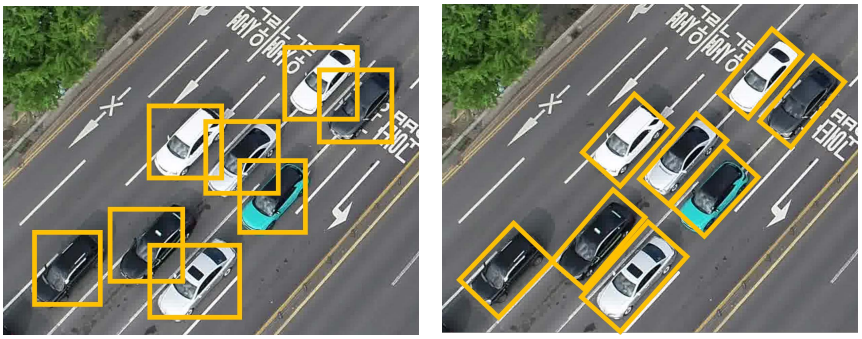


Fig. 3: Visualization of original annotation method(left) and proposed annotation method of DOTA(right)

3 Proposed method

The base-line network of proposed method in this paper is YOLOv5 [12] model. We propose the Dilated Depth-wise block for more efficient calculation and to get result faster instead of C3 module.

3.1 Detection strategy

As mentioned above, there are three kinds of challenges in object detection tasks with drone imagery. The common point of those problems is high resolution. High resolution means that the size of the image is large, in other words, it means that a large area can be observed at once. Convolution operation in a wide area naturally increases the amount of computation and slows down the speed of learning and deriving results. In this paper, we study how to speed up learning and result derivation by reducing the amount of computation without reducing performance. In addition, one of the characteristics of a high resolution is that various types and a large number of objects appear simultaneously. So, at the end of the network, the detection head is increased to detect more objects in the high resolution drone image.
Therefore, the detection strategy of this paper focuses on two things mentioned above. First, The proposed network focuses on speeding up by reducing the amount of computation without reducing the detection accuracy of high resolution images. It can be solved using dilated convolution and depth-wise convolution. The dilated convolution calculates the same number of pixels as normal convolution, but it can calculate a wide area by widening the receptive field. At the same time, In order to reduce the number of calculation parameters, the calculation process is going on for each channel like depthwise convolution. Second, We increase the detection head from three scales to four scales to detect more objects in drone images. Drone image includes small-sized objects like a person, medium-sized objects like cars and trucks, large sized objects like trees and streetlamps, and extra large-sized objects like apartments and buildings as shown in Figure 2. Therefore, detection heads of at least four scales must exist to increase the probability of detecting many objects of various sizes.

3.2 Proposed module

The network proposed in this paper uses the Dilated Depth-wise block introduced in Figure 4 instead of the c3 module used in the YOLOv5 backbone. The c3 module performs operations using only 1x1 convolutions. However, the receptive field is not considered at all when adopting c3 module in original YOLOv5. Therefore, the proposed module uses a mixture of normal convolution and dilated convolution properly. In addition, each convolution layer adopts a depth-wise convolution method to improve detection speed by reducing the amount of computation parameter.
The flow of the Dilated Depth-wise block is as follows. When the feature map is input, the feature map channel is divided into 4 parts without any calculation process. This is because there are other methods such as point-wise convolution, but the use of convolution layers directly increases the amount of computation. The feature maps divided into 4 parts pass into different calculation methods like (1) depth-wise convolution with size 1x1, (2) depth-wise convolution with size 3x3, (3) depth-wise convolution with a dilation ratio of 2, (4) depth-wise convolution with a dilation ratio of 3. The sum of the number of channels of the

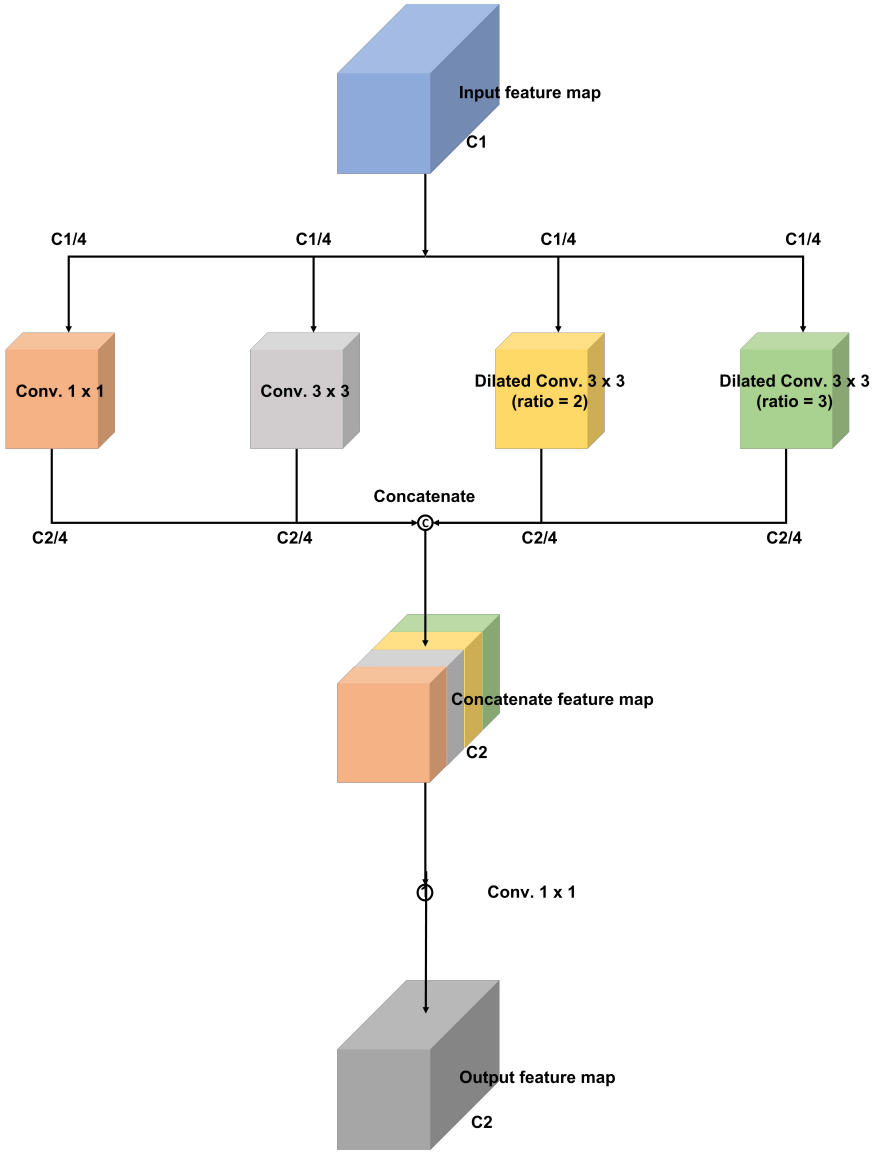


Fig. 4: Illustration of the proposed Dilated Depth-wise module.

feature map generated through a total of 4 operations (1 to 4) is equal to the number of channels in the output feature map. The four feature maps generated are concatenated and passed to a 1x1 convolutional layer. The reason why the concatenated feature map is subjected to 1x1 convolution operation is that information in a narrow area and information in a wide area exist without sharing each other, so that information is shared using 1x1 convolution with the least

amount of computation to obtain higher accuracy.
Also, the number of detection heads increased from 3 to 4. As mentioned in the detection strategy, object scales in drone images can be divided into four scales (small, medium, large, and extra-large). Therefore, the number of scales of the detection head was modified according to the number of object scales.

4 Experiment

4.1 Dataset

The data used in the experiment are the autonomous drone dataset [20] built by University of Ulsan and the VisDrone dataset [21] built by Tianjin University. The challenge and proposed method of drone object detection presented in this paper are from the analysis of the autonomous drone dataset. After network modification, the experiments of same conditions were applied to the VisDrone dataset to prove that proposed network is not overfitted at autonomous drone dataset built by University of Ulsan.

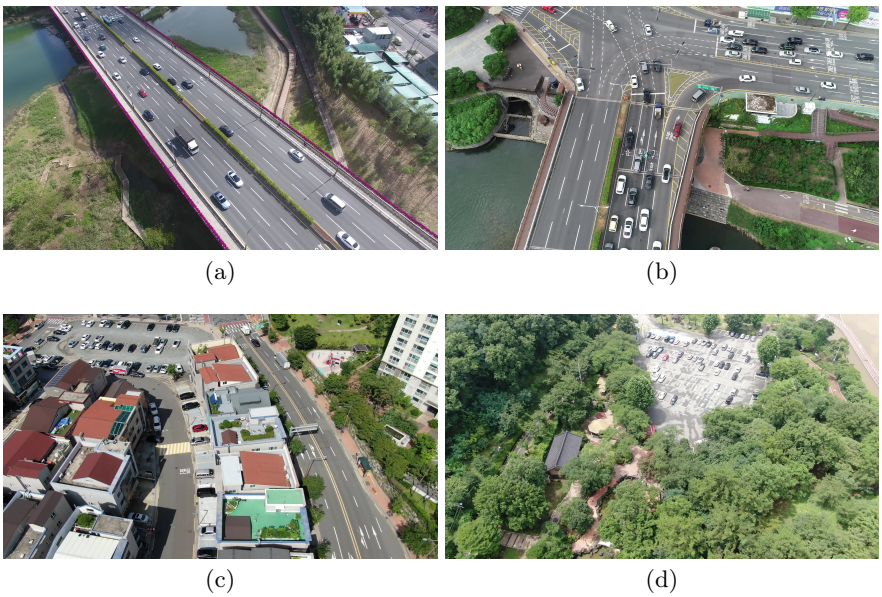
The autonomous drone dataset provides videos, images, and JSON files taken at various altitudes and angles in tourist areas, city areas, and forest areas. Among them, tourist areas and city areas data were mainly used to build a similar environment for future work as we mentioned above such as traffic analysis and surveillance systems. The information of data used in this paper is shown in Table 1 and Figure 5.

Table 1: The information of data used in the experiment.

Category	Region_Place	Altitude	Angle	The number of image
City	Ulsan_Taehwa-bridge	70m	60°	2,343
City	Ulsan_Samho-bridge	60m	45°	2,291
City	Daegu_Geumho-district	60m	45°	1,854
Tourist	Daegu_Hwawon-amusement-park	80m	45°	1,768

4.2 Evaluation metrics

To evaluate the performance of the proposed network, accuracy and speed were used as evaluation criteria. In the case of accuracy of the network, two indicators are measured: mAP 0.5 and mAP 0.5 to 0.95. mAP (mean average precision) represents the average of the area under the PR curve for each class and is a metric for analyzing object detection accuracy through performance evaluation of precision and recall. The number after mAP means IoU (Intersection over Union) value. It means the value measured when the IoU score is 0.5 and measured when IoU score is gradually increasing the value from 0.5 to 0.95. In the case of network speed, parameters, GFLOPS (GPU FLoating point Operations Per Second), and calculation speed per image were used as indicators for validation.



(a) (b)

(c) (d)

Fig. 5: Example images for each region of the autonomous drone dataset, (a) Ulsan_Samho-bridge, (b) Ulsan_Taehwa-bridge, (c) Daegu_Geumho-district, (d) Daegu_Hwawon-amusement-park.

4.3 Implementation setup

All experiments were conducted in the same environment, and the configuration environment was Intel Core i9-9960X, NVIDIA RTX 2080 Ti x 4EA, 125.5 GB memory. The training process was 100 epochs in all experiments, and all hyper-parameters such as batch size and learning rate, and depth multiple were set the same.

4.4 Result

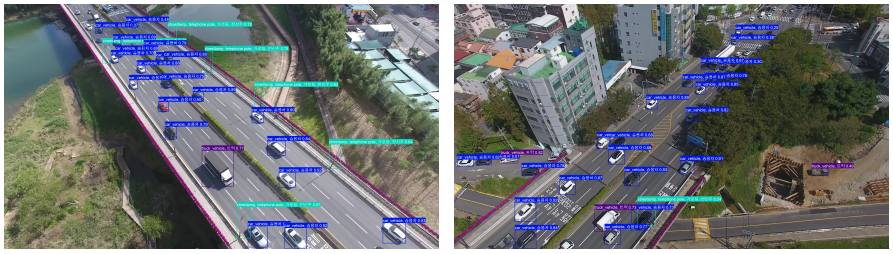
The experimental results are detailed in Tables 2 and 3. The result applied to the autonomous drone dataset is shown in Table 2, and Table 3 is the result applied to the VisDrone dataset. Both results reduced the number of parameters and GFLOPS by about 20%. But, the object detection accuracy was slightly decreased. In the case of autonomous drone dataset, mAP50 decreased by 2.3% and mAP50-95 by 5.5%. Although the accuracy was decreased, as you can see in Figure 6, it can be seen that most objects in the image are well detected. In the case of VisDrone dataset, mAP50 decreased by 0.6%, and mAP50-95 increased by 0.23%. The biggest feature is object detection time. The original network recorded a total of 12.7 ms for pre-processing, inference, and NMS per image. However, the proposed network completed the same process in 6.5 ms per image.

Table 2: The result comparison between DDConv and C3 module using autonomous drone dataset.

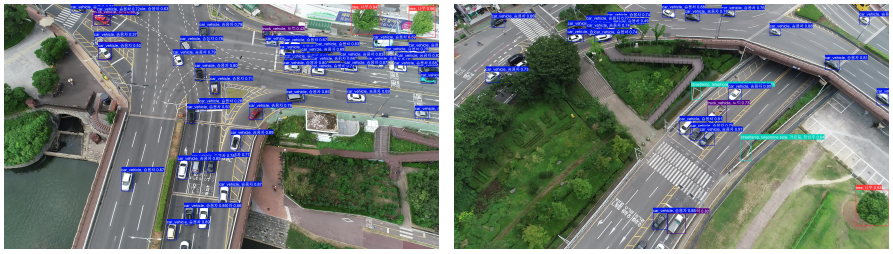
Autonomous drone	DDConv		C3		Performance	
mAP(%)	50	50-95	50	50-95	50	50-95
all	60.3	38.8	62.6	44.3	2.3↓	5.5↓
tree	92.8	67.4	94.1	73.1	1.3↓	5.7↓
person	3.6	0.5	0.02	0	**3.58** ↑	**0.5** ↑
house	89.2	70.4	93.1	78.2	3.9↓	7.8↓
apartment	90.4	62	94.5	69.1	4.1↓	7.1↓
traffic sign	63.7	27.2	75.5	36.1	11.8↓	8.9↓
traffic light	15.7	4.3	0	0	**15.7** ↑	**4.3** ↑
streetlamp	78.8	41.1	84.2	51.7	5.4↓	10.6↓
car	92.9	62.3	94	68.3	1.1↓	6.0↓
bus	66.5	50.2	73.5	59.9	7.0↓	9.7↓
truck	69.6	41.4	75.9	51.5	6.3↓	10.1↓
Parameters	1,437,924		1,783,519		**19.38**% ↓	
GFLOPS	3.3		4.2		**21.43**% ↓	

Table 3: The result comparison between DDConv and C3 module using VisDrone dataset.

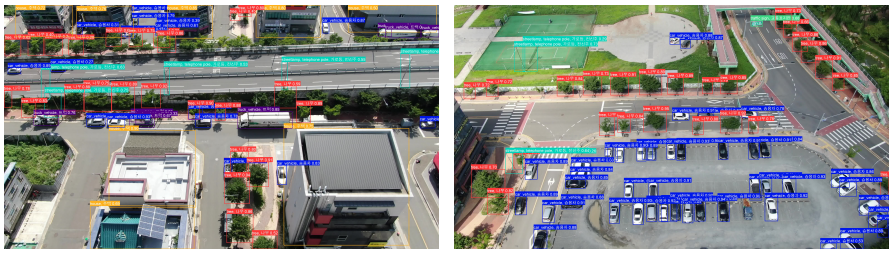
VisDrone	DDConv		C3		Performance	
mAP(%)	50	50-95	50	50-95	50	50-95
all	18.9	9.46	19.5	9.23	0.6↓	**0.23** ↑
pedestrian	21.24	7.37	26.2	9.53	4.96↓	2.16↓
people	17.2	5.54	22.9	7.31	5.7↓	1.77↓
bicycle	2.34	0.74	2.22	0.76	**0.12** ↑	0.02↓
car	54.6	32.4	56.6	33.7	2.0↓	1.3↓
van	18.8	12.0	14.1	8.77	**4.7** ↑	**3.23** ↑
truck	14.2	7.75	13.0	6.69	**1.2** ↑	**1.06** ↑
tricycle	9.21	4.42	8.98	4.28	**0.23** ↑	**0.14** ↑
awning-tricycle	4.91	2.85	3.7	2.06	**1.21** ↑	**0.79** ↑
bus	24.7	14.1	20.3	9.85	**4.4** ↑	**9.85** ↑
motor	21.8	7.43	27.5	9.37	5.7↓	1.94↓
Parameters	1,424,004		1,772.695		**19.67**% ↓	
GFLOPS	3.3		4.2		**21.43**% ↓	



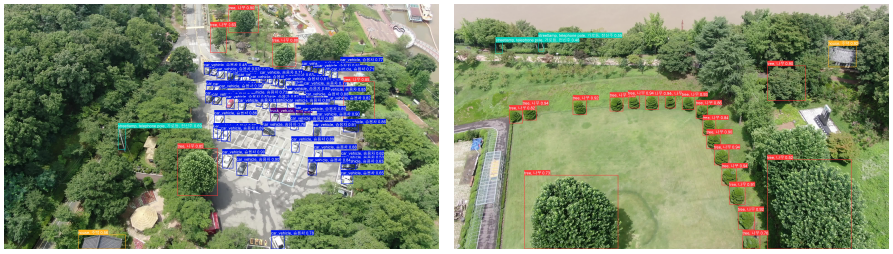
(a) Ulsan_Samho-bridge



(b) Ulsan_Taehwa-bridge



(c) Daegu_Geumho-district



(d) Daegu_Hwawon-amuesment-park

Fig. 6: Visualization of detection results on autonomous drone dataset using YOLOv5 with DDConv block.

5 Conclusion

This paper focused on the object detection work that is the basis of the technologies used in drone-based artificial intelligence systems. So, We proposed a DDConv to reduce the amount of computation of an object detection network for drone systems in which real-time is important. The DDConv includes dilated convolution and depth-wise convolution together to analyze a large area efficiently and to reduce the amount of computation for real-time systems. In addition, a detection head was added at the end of the network to find objects of more diverse scales. As a result of the experiment on the autonomous drone dataset, mAP50 decreased by 2.3% and mAP50-95 by 5.5%. In the case of the VisDrone dataset, mAP50 decreased by 0.6%, and mAP50-95 increased by 0.23%. But, both of two experiments decreased parameters and GFLOPS by about 20%. The object detection speed is almost twice as fast as the original network. The proposed network spends only 6.5 ms per image for inference. Although the accuracy is slightly lower than the original network, the majority of objects are detected as shown in Figure 6. Therefore, the proposed network in this paper is suitable for a drone-based real-time object detection systems.

6 Acknowledgements

This results was supported by "Regional Innovation Strategy (RIS)" through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(MOE)(2021RIS-003)

References

1. C. Chen, H. Min, Y. Peng, Y. Yang, and Z. Wang, "An intelligent real-time object detection system on drones," *Applied Sciences*, vol. 12, no. 20, 2022. [Online]. Available: https://www.mdpi.com/2076-3417/12/20/10227
2. J. Lee, J. Wang, D. Crandall, S. Šabanović, and G. Fox, "Real-time, cloud-based object detection for unmanned aerial vehicles," in *2017 First IEEE International Conference on Robotic Computing (IRC)*, 2017, pp. 36–43.
3. J. Choi and K. Jo, "Lightweight bird eye view detection network with bridge block based on yolov5," in *2022 International Workshop on Intelligent Systems (IWIS)*, 2022, pp. 1–4.
4. Y. Zhang, C. Wang, X. Wang, W. Zeng, and W. Liu, "FairMOT: On the fairness of detection and re-identification in multiple object tracking," *International Journal of Computer Vision*, vol. 129, no. 11, pp. 3069–3087, sep 2021. [Online]. Available: https://doi.org/10.1007%2Fs11263-021-01513-4
5. S. Liu, X. Li, H. Lu, and Y. He, "Multi-object tracking meets moving uav," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 8866–8875.
6. W. Huang, X. Zhou, M. Dong, and H. Xu, "Multiple objects tracking in the uav system based on hierarchical deep high-resolution network," *Multimedia Tools Appl.*, vol. 80, no. 9, p. 13911–13929, apr 2021. [Online]. Available: https://doi.org/10.1007/s11042-020-10427-1

7. Y. Lin, M. Wang, W. Chen, W. Gao, L. Li, and Y. Liu, "Multiple object tracking of drone videos by a temporal-association network with separated-tasks structure," *Remote Sensing*, vol. 14, no. 16, 2022. [Online]. Available: https://www.mdpi.com/2072-4292/14/16/3862
8. X. Wu, W. Li, D. Hong, R. Tao, and Q. Du, "Deep learning for unmanned aerial vehicle-based object detection and tracking: A survey," *IEEE Geoscience and Remote Sensing Magazine*, vol. 10, no. 1, pp. 91–124, mar 2022. [Online]. Available: https://doi.org/10.1109%2Fmgrs.2021.3115137
9. X. Zhang, E. Izquierdo, and K. Chandramouli, "Dense and small object detection in uav vision based on cascade network," in *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, 2019, pp. 118–126.
10. H. Zhou, A. Ma, Y. Niu, and Z. Ma, "Small-object detection for uav-based images using a distance metric method," *Drones*, vol. 6, no. 10, 2022. [Online]. Available: https://www.mdpi.com/2504-446X/6/10/308
11. M. Liu, X. Wang, A. Zhou, X. Fu, Y. Ma, and C. Piao, "Uav-yolo: Small object detection on unmanned aerial vehicle perspective," *Sensors*, vol. 20, no. 8, 2020. [Online]. Available: https://www.mdpi.com/1424-8220/20/8/2238
12. G. Jocher, A. Chaurasia, A. Stoken, J. Borovec, NanoCode012, Y. Kwon, K. Michael, TaoXie, J. Fang, imyhxy, Lorna, Yifu), C. Wong, A. V, D. Montes, Z. Wang, C. Fati, J. Nadar, Laughing, UnglvKitDe, V. Sonck, tkianai, yxNONG, P. Skalski, A. Hogan, D. Nair, M. Strobel, and M. Jain, "ultralytics/yolov5: v7.0 - YOLOv5 SOTA Realtime Instance Segmentation," Nov. 2022. [Online]. Available: https://doi.org/10.5281/zenodo.7347926
13. F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," 2015. [Online]. Available: https://arxiv.org/abs/1511.07122
14. T.-Y. Zhang, J. Li, J. Chai, Z.-Q. Zhao, and W.-D. Tian, "Improved yolov5 network with attention and context for small object detection," in *Intelligent Computing Methodologies*, D.-S. Huang, K.-H. Jo, J. Jing, P. Premaratne, V. Bevilacqua, and A. Hussain, Eds. Cham: Springer International Publishing, 2022, pp. 341–352.
15. A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.
16. G.-S. Xia, X. Bai, J. Ding, Z. Zhu, S. Belongie, J. Luo, M. Datcu, M. Pelillo, and L. Zhang, "Dota: A large-scale dataset for object detection in aerial images," 2017. [Online]. Available: https://arxiv.org/abs/1711.10398
17. J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," 2015. [Online]. Available: https://arxiv.org/abs/1506.02640
18. J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," 2018. [Online]. Available: https://arxiv.org/abs/1804.02767
19. R. Hamaguchi, A. Fujita, K. Nemoto, T. Imaizumi, and S. Hikosaka, "Effective use of dilated convolutions for segmenting small object instances in remote sensing imagery," 2017. [Online]. Available: https://arxiv.org/abs/1709.00179
20. K. Jo. (2020) Autonomous drone dataset. [Online]. Available: https://aihub.or.kr/aihubdata/data/view.do?currMenu=115topMenu=100
21. P. Zhu, L. Wen, D. Du, X. Bian, H. Fan, Q. Hu, and H. Ling, "Detection and tracking meet drones challenge," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2021.