Small Object Detection without Attention for Aerial Surveillance

Jehwan Choi[0009−0005−8494−2170], Duy-Linh Nguyen[0000−0001−6184−4133], Xuan-Thuy Vo[0000−0002−7411−0697], and Kang-Hyun Jo[0000−0001−4937−7082]

Intelligent Systems Laboratory,
Department of Electrical, Electronic and Computer Engineering,
University of Ulsan, Ulsan 44610, South Korea
cjh1897@ulsan.ac.kr, ndlinh301@mail.ulsan.ac.kr,
xthuy@islab.ulsan.ac.kr, acejo@ulsan.ac.kr

Abstract. This paper introduces the development of a essential deep learning model for surveillance systems utilizing high-mounted CCTV or drones. Objects seen from elevated angles often look smaller and may appear at different angles compared to ground-level observations. To improve the detection of small objects, we propose a network incorporating an element-wise multiplication module based on the vanilla Vision Transformer (ViT) architecture [1]. However, traditional transformer models need significant computational resources, which may not be practical for edge devices like CCTV cameras or drones. Therefore, we apply the Attention-Free Transformer (AFT) [2] to reduce computational requirements enabling real-time operation on low-capacity devices. We validate the performance by combining ViT and AFT with the YOLOv5 real-time object detection model. Practical applicability is confirmed by implementing it on the low-capacity device named ODROID H3+. Validation datasets include Autonomous Driving Drone [3], VisDrone [4], Aerial-Maritime [5], and PKLot [6], all containing numerous small-sized objects. Experimental results on VisDrone dataset show that YOLOv5 [7] nano + AFT reduces parameter count by 4.6% while increasing accuracy by 1%, making it an efficient network. The model size is suitable for edge device implementation at 3.7Mb. Similarly, Aerial Maritime and PKLot datasets indicatge decreased amount of parameters and increased accuracy. Hence, the proposed deep learning model is applicable for aerial surveillance systems.

Keywords: Small object detection · attention-free transformer · edge device · aerial surveillance system.

1 Introduction

In recent years, surveillance systems utilizing cameras mounted on drones and aerial platforms, as well as CCTV, have significantly advanced. Primarily, images captured from a bird's eye view (BEV) offer the advantage of a wide surveillance range, allowing the observation of many objects simultaneously. This presents

new possibilities for various systems including urban monitoring, traffic management, environmental observation, and emergency response. However, such systems have the challenge of the small size of objects within the captured images. Additionally, the complex background makes it difficult to identify small-sized objects. This paper proposes a deep learning model that can effectively detect objects when providing surveillance systems and various services using cameras mounted on drones and aerial platforms.

Object detection is a fundamental task in computer vision and artificial intelligence. Many researchers have implemented real-time object detection systems using the YOLO family [7–10], a Convolutional Neural Network (CNN)-based deep learning model. However, there has been a significant shift from CNN-based architectures to Transformer-based structures [1] in recent deep learning models, achieving remarkable improvements in object detection performance. The main drawback of Transformers is their high computational cost and memory demand. For mobile entities like drones, the computational cost and processing power required pose considerable challenges, making the computational cost of Transformers inappropriate. Similarly, edge devices suitable for mounting on mobile entities have limited processing capabilities, making it impractical to operate surveillance systems using Transformer-based deep learning models. To address this, this paper applies the Attention-Free Transformer (AFT), which has a similar structure to the Vision Transformer (ViT) but reduces computational costs by using element-wise multiplication instead of matrix multiplication. Additionally, AFT enhances the detection performance for small objects by comparing the characteristics of each part of an image, rather than comparing every part with each other, simplifying the feature extraction process and calculating the feature map under conditions favorable for small object detection. Through this, we propose an effective deep learning model for surveillance systems utilizing drones and aerial platforms.

To validate the efficiency of the proposed deep learning model, its performance is evaluated on edge devices. The equipment used in this study is the ODROID H3+, and the model's accuracy and real-time performance are tested using BEV datasets such as VisDrone, xView, PKLot, and AerialMaritime. The main contributions of this paper are outlined as follows:

- Proposal of a deep learning model applicable to aerial surveillance systems, validated through experiments on datasets containing small objects such as VisDrone, xView, PKLot, and AerialMaritime.
- Application of the Attention-Free Transformer (AFT) to reduce computational costs and memory requirements in comparison to the Vision Transformer (ViT).
- Validation of the proposed deep learning model's suitability for edge devices through implementation and testing on the ODROID H3+.

2 Related Work

2.1 Attention-Free Transformer

Since the introduction of the Vision Transformer (ViT), a myriad of networks featuring various architectures and methodologies based on the Transformer paradigm have been proposed like Swin Transformer and EfficientViT. However, a commonality among most Transformer architectures is the inclusion of the core module known as Multi-Head Attention (MHA). As depicted in Figure x (a), MHA operates by segregating the input image into Query (Q), Key (K), and Value (V) components. It then computes the relationship between Q and K through dot products, ultimately enhancing performance by calculating self-relevance through a weighted sum with V. This process is formalized in Equation (1). Because of the MHA use matrix multiplication, allowing the computational complexity to be denoted as such $O(N^2D)$. N and D represent the size and dimension of the feature map input into the transformer, respectively.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

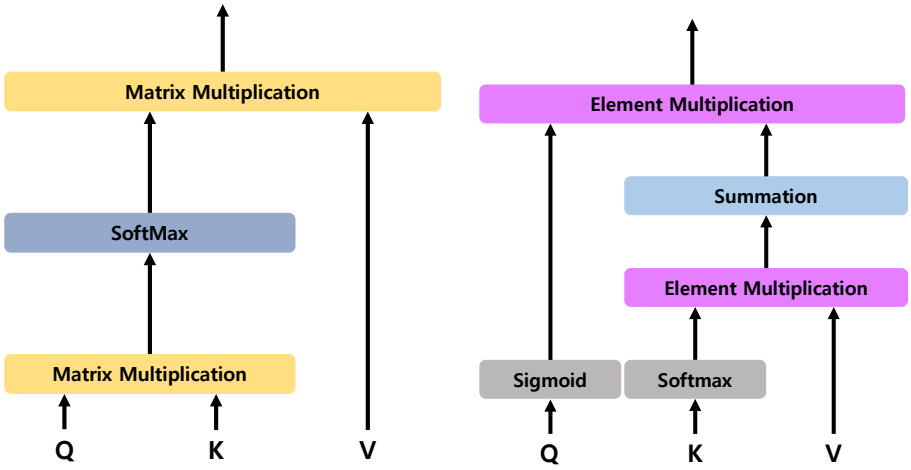
Due to the rapid increase in computational demand as N grows, this paper introduces the Attention-Free Transformer (AFT). Unlike Multi-Head Attention (MHA) that segments input images into Q, K, and V using dot products, AFT employs element-wise multiplication, as shown in Figure x(b). This method changes the operation sequence slightly. AFT calculates a weighted average of value by first applying softmax to K, then combines it with V, and finally assesses the relationship between Q and V, similar to MHA. As a result, AFT's computational complexity is noted as $O(ND)$, which is N times less than MHA, making it more efficient. The process of AFT is detailed in Equation (2).

$$\text{AFT}(Q, K, V) = \sigma(Q) \odot \sum(\text{softmax}(K) \odot V) \tag{2}$$

2.2 Edge Computing

Edge Computing (EC) is establishing itself as a pivotal technology to overcome the limitations faced by Cloud Computing (CC). The principal challenges of CC include latency in data transmission to servers, high costs, and network congestion during large data transfers, as well as cyber security threats. In contrast, EC offers real-time responsiveness by enabling immediate processing at the point of data generation. It enhances system efficiency through distributed data processing and bolsters security with on-site data handling.

Recent research in the field of computer vision underscores the significant role of EC. Guanchu Wang et al. (2022) [11] developed an object detection system for Edge Devices (BED), showcasing an end-to-end pipeline utilizing a compact 300kb Deep Neural Network (DNN) model. Moreover, Shihan Liu et al. (2023) [12] presented a study that achieved processing speeds of over 30 frames



(a) Multi-Head Attetion (MHA) (b) Attention-Free Transformer (AFT)

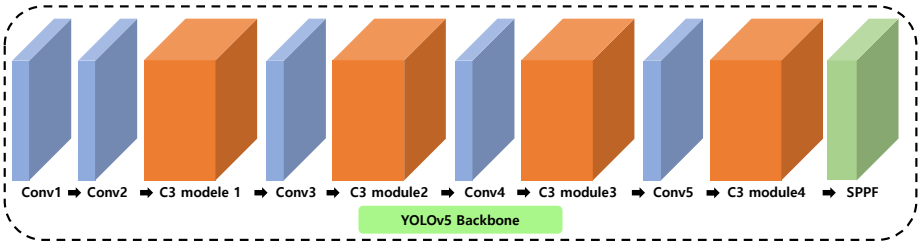
Fig. 1: Architecture of Multi-Head Attention and Attention-Free Transformer

per second on Nvidia Jetson AGX Xavier, employing data augmentation techniques with Mosaic and an Efficient Decoupled Head. These studies illustrate not only how EC surpasses the limitations of CC but also its growing applicability in various domains, including Internet of Things (IoT) devices, smart cities, and autonomous vehicles.

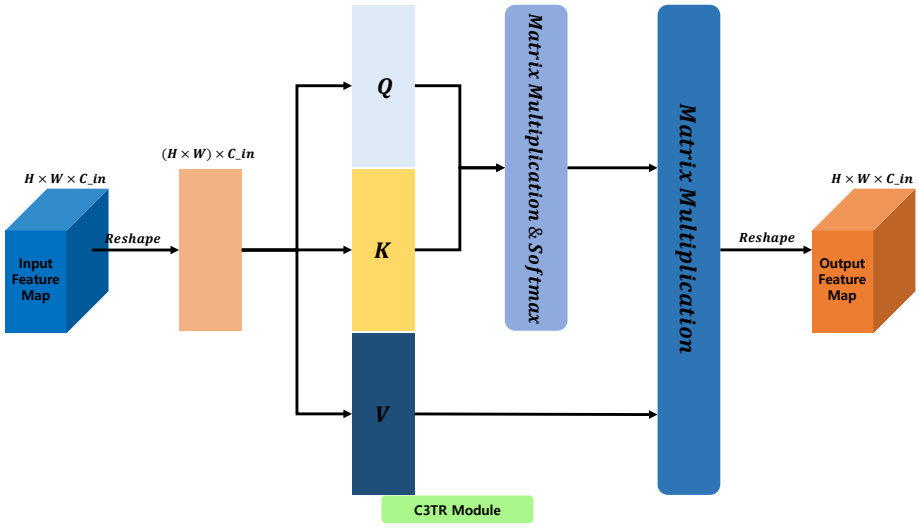
3 Proposed Method

3.1 Overall pipeline

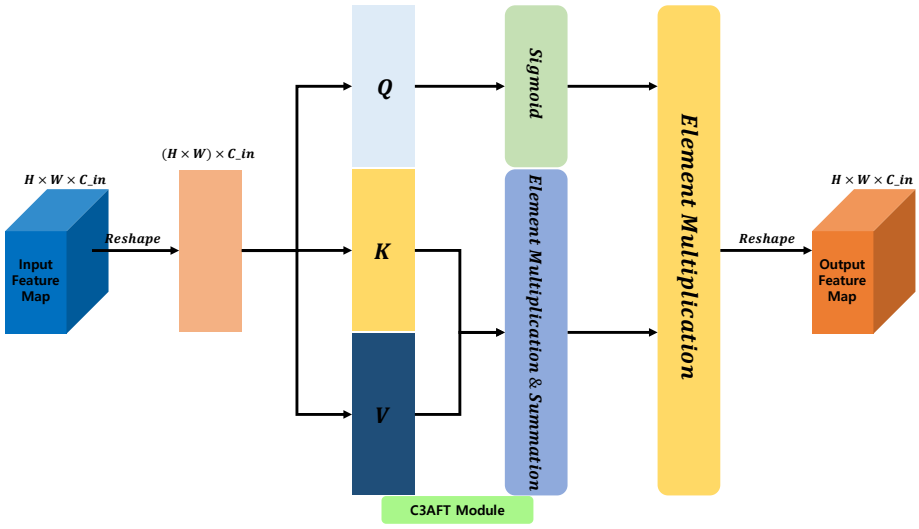
This paper proposes a deep learning model based on YOLOv5, a real-time object detection model with superior performance due to its diverse CNN architectures. The core of YOLOv5 is the Cross Stage Partial (CSP) network strategy and the C3 module which utilizes three convolutional operations. YOLOv5 backbone with four C3 modules is described as shown in Figure 2(a). Replacing the C3 module with C3TR and C3AFT modules for experiments, the C3TR module integrates a transformer structure into the C3 module as shown in Figure 2(b) while the C3AFT module applies AFT to the C3 module illustrated in Figure 2(c). Given transformer structures like C3TR are usually applied towards the network's end, C3AFT was also placed in a similar position to ensure consistent experimental conditions. By substituting the last two parts of the C3 module in the backbone with C3TR and C3AFT modules, the study assesses object detection accuracy and computational speed.



(a) YOLOv5 backbone



(b) C3TR module



(c) C3AFT module

Fig. 2: The overall pipeline of the YOLOv5 backbone (a) and the process of the C3TR (C3 module with transformer) module (b) and C3AFT (C3 module with attention-free transformer) module (c) to replace the C3 module.

3.2 C3AFT module

The C3AFT module is proposed to address the limitations of transformers. As described in Section 2.1, "Attention-Free Transformer," the difference between the Multi-Head Attention (MHA) used in Vision Transformer (ViT) and the AFT applied in this paper lies in the method of operating across feature maps. While MHA determines the important parts of a feature map through matrix multiplication, AFT does so using element-wise multiplication. Particularly, in this paper, AFT is applied to detect small objects within images. There are two main reasons why AFT is advantageous for small object detection. First, by performing operations across the entire feature map, AFT can recognize even relatively small areas occupied by small objects as containing important information. Compared to traditional CNNs, this allows for the extraction of features not just based on local information but also considering the overall context. Second, AFT normalizes the importance of each feature across the entire feature map using the softmax function on the key vector, creating weights, and then combines this result with the value (original feature map) as shown in Figure 2(c). This approach enables better capture of small object characteristics. Finally, by applying a sigmoid operation to the processed query and combining it with the key-value output, important information is enhanced while non-critical information is suppressed, thus extracting significant features of objects.

3.3 Training Strategy

As shown in Figure 2(a), YOLOv5 backbone incorporates the C3 module a total of four times. In this paper, the last two of the four C3 modules are replaced with the C3TR and C3AFT modules for experimentation. Given that transformers are primarily applied in the network's final layers, this paper similarly replaces the last C3 module with the proposed modules. Additionally, to explore the impact of repeating transformer structures on performance enhancement, the third and fourth C3 modules are substituted with the C3TR and C3AFT modules.

4 Experiment

4.1 Dataset

Autonomous Driving Drone: This dataset was created by the Intelligent Systems Laboratory (ISLab), University of Ulsan, Korea, to which I am affiliated, and I, as an author of this paper, participated in its production. The dataset comprises 4k videos shot at various altitudes and angles across tourist spots, urban areas, and forests in Korea. It is categorized into object detection, segmentation, and 3D LiDAR datasets, with the object detection dataset containing over 30 million images and encompassing 18 classes (such as person, tree, house, car, bus, traffic light, etc.). For this paper, data from two tourist locations and two urban areas were used, totaling 10,321 images for training (8,256) and evaluation (2,065).

VisDrone: This dataset was produced by the AISKYEYE team at the Lab of Machine Learning and Data Mining, Tianjin University, China. It consists of over 260,000 high-resolution images taken in various urban, rural, and coastal areas, organized into 10 classes (including pedestrian, car, bicycle, etc.). A total of 8,629 images from this dataset were utilized in this paper.
xView: The xView dataset, constructed by the Defense Innovation Unit Experimental (DIUx) and the National Geospatial-Intelligence Agency (NGA), stands as the largest publicly available satellite imagery dataset. Comprising 1,127 images with over a million objects across 60 classes, it presents challenges for object detection due to its 0.3m resolution. In this paper, 846 images were utilized for training purposes and 281 images for evaluation.
PKLot: The PKLot dataset, produced by the Federal University of Paraná, Brazil, encompasses 12,416 images captured by parking lot surveillance cameras. Designed to determine the presence or absence of vehicles in parking areas, this dataset includes images taken under various weather conditions, such as sunny, cloudy, and rainy. For the purposes of this paper, 9,933 images were employed for training, and 2,483 images were used for evaluation.
AerialMaritime: Constructed by the team led by Jacob Solawetz at Roboflow, this dataset comprises data captured over maritime environments using a Mavic Air 2, a compact drone. Shot in 4k resolution from an altitude of 400ft (approximately 122 meters), it includes a total of 508 images categorized into five classes. In this paper, 393 images were utilized for training purposes and 105 images for evaluation.

4.2 Evaluation Metric

In the evaluation framework, focus lies on analyzing model efficiency and performance through key metrics. First, model complexity gets examined by measuring the total number of parameters, directly influencing computational efficiency. Furthermore, model processing capability in terms of FLOPS (Floating Point Operations Per Second) undergoes evaluation, offering insights into computational speed.
Additionally, mAP50 and mAP50-95 serve as primary accuracy metrics for a comprehensive assessment of model precision in object detection. These indicators assist in gauging model effectiveness across various detection thresholds. Lastly, model size in megabytes (Mb) and inference speed are considered for model feasibility on Edge Devices. This ensures meeting operational requirements for edge computing environments.

5 Result

5.1 Quantitative Result

The results from the five datasets mentioned earlier are shown in Table 1. For all datasets, the original YOLOv5 had the fewest layers. The proposed model's layer

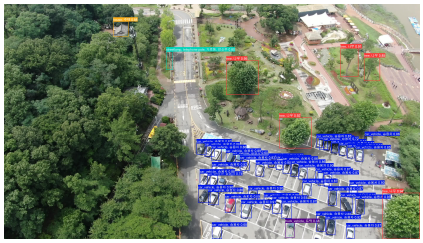
count differed by only 1 to 3 layers compared to the best results, indicating similar performance since layer count does not directly affect model lightness. The proposed model showed the best results in parameter count across all datasets, leading to the smallest model size due to having the fewest parameters among the five compared networks. The best performance in GFLOPS was observed when the C3TR module was used twice, likely because the C3AFT module, while requiring less computation, has a more complex structure. In terms of mAP50, modules applying AFT generally achieved the best results. The Autonomous Driving Drone dataset showed nearly identical accuracy to the best result, being only 0.1% lower. For mAP50-95, networks with the proposed module performed best on the VisDrone and Aerial Maritime datasets. In the Autonomous Driving Drone and PKLot datasets, the proposed model recorded 0.2% and 0.8% lower scores than the best results, respectively, which is less than a 1% difference, thus considered equivalent in performance.
However, using the C3TR module twice yielded the best performance in GFLOPS, suggesting that the C3AFT module, despite having lower computational requirements, has a more complex structure. In terms of mAP50, modules applying AFT generally achieved the best results. The Autonomous Driving Drone dataset recorded a 0.1% lower score compared to the best result, indicating nearly identical accuracy. For mAP50-95 scores, networks with the proposed module performed best in the VisDrone and Aerial Maritime datasets. In the Autonomous Driving Drone and PKLot datasets, they scored 0.2% and 0.8% less than the best results, respectively, but these differences are under 1%, so the performances are deemed equivalent.

5.2 Qualitative Result

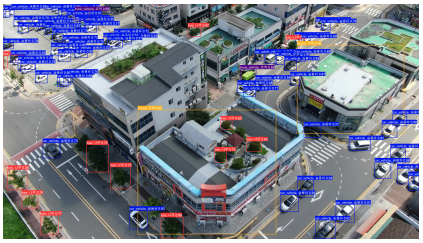
The result images were extracted from models trained on each dataset using an ODROID H3+. The xView dataset was excluded because the detection accuracy is significantly low. In the Autonomous Driving Drone data, large objects like buildings and small objects like vehicles, trees, and streetlamps are almost detected. Similarly, the VisDrone data showed good results in detecting vehicles and pedestrians. In the case of PKLot, the results were concentrated in parking spaces located at the center of the images, indicating the labeling focus on central areas. This problem means that data augmentation and diversity in datasets are very important. In conclusion, the proposed network in this paper has been demonstrated to effectively detect various objects including small objects with high accuracy.

5.3 Ablation Study

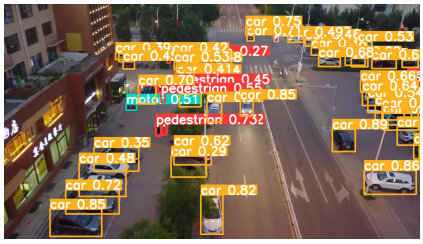
This paper primarily uses datasets captured by drones or aerial platforms featuring small objects. For the Autonomous Driving Drone data, humans measure about 20-30 pixels and vehicles 100-150 pixels against a total image size of 4k (3840 x 2160), indicating the small size of objects. Therefore, this study tests the proposed network's performance on datasets with larger objects, using the



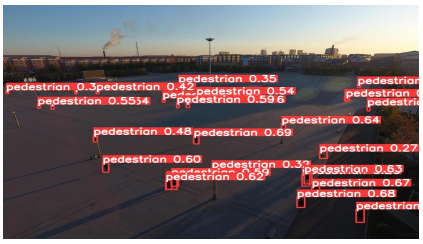
(a) Autonomous Driving Drone result 1



(b) Autonomous Driving Drone result 2



(c) VisDrone result 1



(d) VisDrone result 2

(e) PKLot result 1

(f) PKLot result 2

(g) AerialMaritime result 1

(h) AerialMaritime result 2

Fig. 3: The result images of proposed method applied to aerial datasets.

Table 1: Comparison of five networks on datasets containing small objects.

Dataset	Architecture	Layers	Parameters	GFLOPS	mAP50	mAP50-95	Model size
Autonomous Driving Drone	Original	**214**	1,788,271	4.3	**63.9**	**45.1**	3.9
	C3TR	217	1,788,399	4.2	63.4	44.9	3.9
	C3TR*2	222	1,780,271	**4.1**	62.6	43.2	3.8
	C3AFT	215	**1,706,223**	4.2	63.8	44.9	**3.7**
	C3AFT*2	241	1,711,151	4.2	62.7	43.4	**3.7**
VisDrone	Original	**157**	1,772,695	4.2	19.8	9.45	3.8
	C3TR	162	1,773,079	4.1	19.6	9.31	3.8
	C3TR*2	173	1,765,335	**3.9**	17.7	8.54	3.8
	C3AFT	160	**1,690,903**	4.1	**20.8**	**9.94**	**3.7**
	C3AFT*2	186	1,695,703	4.1	17.8	8.41	**3.7**
Aerial Maritime	Original	**157**	1,765,930	4.1	36.9	17.8	3.9
	C3TR	162	1,766,314	4.1	38.5	18.9	3.9
	C3TR*2	173	1,758,570	**3.9**	33.1	16.2	3.8
	C3AFT	160	**1,684,138**	4.1	**42.2**	**20**	**3.7**
	C3AFT*2	186	1,688,938	4.1	36.4	15.4	**3.7**
PKLot	Original	**157**	1,761,871	4.1	**99.5**	**92.1**	3.9
	C3TR	162	1,762,255	4.1	99.4	91.5	3.9
	C3TR*2	173	1,754,511	**3.9**	99.4	89	3.9
	C3AFT	160	**1,680,079**	4.1	**99.5**	91.3	**3.7**
	C3AFT*2	186	1,684,879	4.1	99.4	89.4	3.8
xView	Original	**157**	1,840,345	4.4	0.72	**0.27**	4.1
	C3TR	162	1,840,729	4.3	0.648	0.482	4.1
	C3TR*2	173	1,832,985	**4.1**	0.305	0.0809	4
	C3AFT	160	**1,758,553**	4.3	**1.24**	0.181	**3.7**
	C3AFT*2	186	1,763,353	4.3	1.04	0.212	3.9

Visual Object Classes (VOC) dataset and comparing results across five networks as shown in Table 1. The experiment results for the VOC data presented in Table 2 show a performance decline in networks applying both C3TR and C3AFT, with a more significant decrease in networks with C3AFT. A roughly 3% accuracy drop suggests the performance slightly worsens on datasets with larger objects. The reason is that AFT applies element-wise multiplication across the entire feature map, which can enhance representation where small objects are present but may reduce expressiveness for larger objects as the entire feature map is highlighted.

Another ablation study involves validating performance on the edge device ODROID H3+. Given the challenging power supply conditions for drones and aerial platforms, leveraging low-power edge devices for data processing is advantageous. Having already concluded that the proposed network's model size is the smallest, performance is further proven by operating it on the edge device and comparing inference times. The comparative results of inference times are shown in Table 3.

Table 2: Comparison result of VOC dataset containing big objects.

Architecture	Layers	Parameters	GFLOPS	mAP50	mAP50-95	Model size
Original	**157**	1,786,225	4.2	**65.6**	**38.4**	3.9
C3TR	162	1,786,609	4.2	64	36.3	3.9
C3TR*2	173	1,778,865	**4**	59.2	31.8	3.9
C3AFT	160	**1,704,433**	4.1	62.9	35.9	**3.7**
C3AFT*2	186	1,709,233	4.2	57.3	30.4	**3.7**

Table 3: Comparison inference time of five networks on the ODROID H3+.

Dataset	Original	C3TR	C3TR*2	C3AFT	C3AFT*2
Autonomous Driving Drone	**179.5**	198.9	306.6	**183.1**	239.1
VisDrone	**130.2**	133.8	234.6	**132.6**	168.1
Aerial Maritime	**153.4**	158.1	328.4	**155.1**	202.6
PKLot	**183.9**	194	485.4	**187.9**	249.5
xView	**410.2**	452	832.4	**425.8**	658.2

In all datasets, the original YOLOv5 demonstrates the fastest inference time, with networks applying C3AFT once showing the second fastest inference times. However, the difference is minimal, ranging from just 1.7ms to 15.6ms per image. Calculated in frames per second (fps), the Autonomous Driving Drone dataset shows the original at 5.57fps, C3TR at 5.03fps, and C3AFT at 5.46 fps, indicating similar computational speeds. Thus, networks utilizing AFT prove to perform well with computational speeds comparable to the original YOLOv5.

6 Conclusion

This study introduces a deep learning model for efficiently detecting objects in images taken from drones and high-positioned CCTVs. Especially, focusing on the challenge of detecting small objects in aerial images. The proposed deep learning model uses CNN and transformer architecture simultaneously to generate low computational feature map and to understand the image's full context and highlight significant parts, respectively. The proposed 'C3AFT' module addresses the transformer's computational demand through element-wise multiplication. It enhances vital information by generating a weight vector for the entire image and interacting with the original feature map to improve performance by identifying essential features including those of small objects. Experimental results with five datasets including numerous small objects show the proposed model reduces parameters by approximately 5% and model size by an average of 0.22Mb compared to the original YOLOv5 nano. The detection accuracy improved on three datasets, remained consistent on one dataset, and slightly

decreased by 0.1% on another. Inference time comparisons on the edge device ODROID H3+ reveal the proposed network is marginally slower by an average of 5.46ms, equating to processing one fewer image every nine seconds, which is considered nearly identical in performance. These findings demonstrate that the YOLOv5 model with applied AFT efficiently processes BEV images better than the original YOLOv5 model, proving its viability for aerial surveillance systems.

References

1. A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," 2021.
2. X. Wang, Z. Liu, Y. Hu, W. Xi, W. Yu, and D. Zou, "Featurebooster: Boosting feature descriptors with a lightweight neural network," 2023.
3. K. Jo. (2020) Autonomous drone dataset. [Online]. Available: https://aihub.or.kr/aihubdata/data/view.do?currMenu=115topMenu=100
4. 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.
5. J. Solawetz, "Aerial maritime drone dataset," 07 2020. [Online]. Available: https://public.roboflow.com/object-detection/aerial-maritime
6. P. Almeida, L. Soares de Oliveira, A. Jr, E. Jr, and A. Koerich, "Pklot - a robust dataset for parking lot classification," *Expert Systems with Applications*, vol. 42, 02 2015.
7. 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
8. 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
9. J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," 2018. [Online]. Available: https://arxiv.org/abs/1804.02767
10. G. Jocher, A. Chaurasia, and J. Qiu, "Ultralytics yolov8," 2023. [Online]. Available: https://github.com/ultralytics/ultralytics
11. G. Wang, Z. P. Bhat, Z. Jiang, Y.-W. Chen, D. Zha, A. C. Reyes, A. Niktash, G. Ulkar, E. Okman, X. Cai, and X. Hu, "Bed: A real-time object detection system for edge devices," 2022.
12. S. Liu, J. Zha, J. Sun, Z. Li, and G. Wang, "Edgeyolo: An edge-real-time object detector," 2023.