2023 IEEE International Conference on Robotics and Automation (ICRA 2023)
May 29 - June 2, 2023. London, UK

CEAFFOD: Cross-Ensemble Attention-based Feature Fusion Architecture Towards a Robust and Real-time UAV-based Object Detection in Complex Scenarios

Ahmed Elhagry[1], Hang Dai[1], Abdulmotaleb El Saddik[1,2], Wail Gueaieb[1,2], Giulia De Masi[3].

***Abstract*— Deploying object detectors in embedded devices such as unmanned aerial vehicles (UAVs) comes with many challenges. This is due to both the UAV itself having low embedded resources in terms of computation and memory, and also due to the nature of the captured visual data with the variations in objects' scale, orientation, density, viewpoint, distribution, shape, context and others. It is crucial for the object detector to be robust with high accuracy, real-time with fast inference and light-weight to be applicable. Inspired by YOLO architecture, we propose a novel single-stage detection architecture. Our contributions are, first, feature fusion spatial pyramid pooling (FFSPP) block that applies attention-based feature fusion across both time and space utilizing the information of subsequent frames and scales in an efficient manner. Secondly, we introduce a multi-dilated attention-based cross-stage partial connection (MDACSP) block that helps in increasing the receptive field and producing per-channel modulation weights after aggregating the feature maps across their spatial domain. Third, scaled feature fusion head (SFFH) fuses both the FFSPP block features and the connected MDACSP block features specific for this head. For a more robust result across different scenarios, we perform cross-ensembling with three of the top UAV/traffic surveillance datasets: UAVDT, UA-DETRAC and VisDrone. Our ablation study shows how every contribution improves over the baseline. Our approach yielded the state-of-the-art results in all the aforementioned datasets achieving 89.3% mAP, 93.5% mAP, and 42.9% mAP respectively. Testing the model performance on NVIDIA Jetson Xavier NX board shows a desirable balance between the inference time and the memory cost. We also show qualitatively the model robustness and efficiency across the diverse complex scenarios of these datasets. We hope this work facilitates the advancement of the UAV-based perception in such crucial industrial applications.**

I. Introduction

Unmanned aerial vehicles (UAVs) offer a greater mobility and a larger field of view compared to stationary cameras [1]. It features flaws such as a shifting background, low resolution, and lighting fluctuations [2]. Object detection in UAV's images or videos, on the other hand, is not the same as standard object detection. In aerial images, not only the scale of object instances changes, object orientation differs arbitrarily [3]. Class imbalance also happens in real life, where a small-sized object like a car exists more than a truck [4]. Moreover, distribution of objects differ according to the location, where a parking space will be far more congested than a highway, and it alternates at intersections and traffic lights [5]. Beside all of that, other factors increase the complexity of the problem. For instance, weather condition where it can be foggy or rainy, or the time being day or night, will affect the lighting condition and the clarity of objects [6]. In addition, the camera view and the flying altitude of the UAV will impact the objects scale and orientation. These factors have also an impact on occlusion, where having the objects partially visible or out-of-the-view happens frequently [7]. In this work, we introduce a novel single-stage detection architecture engineered to balance between the UAV low computational resources and dependable real-time detection accuracy for the complex scenarios of highway surveillance and traffic management. The rest of this paper is organized as follows: Section II covers the literature related to object detectors and single-stage ones. Then it moves to discuss some promising architectures, networks and techniques that had impact on the perception of embedded devices. Section III talks about our approach and different contributions we propose. Section IV investigates the datasets and their associated challenges. It also explains briefly the used evaluation metrics and implementation details of our experiments. Section V discusses the quantitative and qualitative results and how the different components contributed to the overall improvement tackling the UAV-based perception challenges.

Fig. 1. The investigated datasets' challenging shooting conditions [8]–[10]

[1]Ahmed Elhagry, Dr. Hang Dai, Prof. Abdulmotaleb El Saddik and Prof. Wail Gueaieb are with Mohamed Bin Zayed University of Artificial Intelligence (MBZUAI), Masdar City, Abu Dhabi, United Arab Emirates. [2]Prof. Abdulmotaleb El Saddik and Prof. Wail Gueaieb are also with University of Ottawa, Canada. [3]Dr. Giulia De Masi is with Technology Innovation Institute (TII), Masdar City, Abu Dhabi, United Arab Emirates. {Ahmed.Elhagry, Hang.Dai, A.Elsaddik, Wail.Gueaieb}@mbzuai.ac.ae, Giulia.Demasi@tii.ae

II. Related Work

A. Single-stage Object Detectors

Algorithms for detecting objects, such as region-based detectors, excelled in terms of accuracy [25]. The speed rate, nevertheless, is ineffective. Compared to region-based detection systems, single-shot detection methods are fast and memory-light. In order to identify several objects utilizing the multi-box in an image, algorithms that use the single-shot detection approach [24] only need to capture a single shot. Due to the fact that bounding box proposals like those found in RCNN [23] are not employed, it is significantly quicker and has great accuracy. In order to predict the object classes and localize them, it also has a convolution filtering with a steadily diminishing gain.

B. You Only Look Once (YOLO) Architecture

YOLO [26] adopted a fresh approach to object detection. The object localization and feature extraction modules are merged into a single unified entity. Additionally, the classification and localization heads are integrated. Fast inference times are produced by this single-stage pipeline. This innovative approach has gotten the idea of edge devices ever closer to reality when combined with the other detectors based on MobileNet [27]. The idea behind YOLO is that there are not repeated region suggestions as in region-based detectors, nor are there distinct categorization or detection modules that need to be synced with one another. All tasks, including feature extraction, boundary box regression, and classification, are carried out by the one and only single unified structure. There is only one output layer, but it has several properties. Several recent YOLO-based approaches tried adopting the UAV perception problem by focusing on detecting small and distorted objects through attention and augmentations, but their accuracy is still comparatively low to be industrially deployed [32]–[34].

C. Cross Stage Partial Network (CSPNet)

To deploy a computer vision model on an embedded mobile device as the unmanned aerial vehicle, many constraints exist in terms of computation and memory resources [13]. Getting high accuracy in tasks such as object detection is highly desirable as well in many applications. It becomes even crucial in applications such as traffic and crowd management, surveillance and security. The related issue that earlier works need intensive inference calculations from a network design perspective is mitigated by CSPNet [15]. By incorporating feature maps from the start and end of a network stage, the network accounts for gradient variations in an efficient way [14].

D. Spatial Pyramid Pooling (SPP)

Prior to the spatial pyramid pooling [28], the extracted feature map was often flattened into a one-dimensional vector, then applied in a sliding window method, producing an output of varying sizes. The CNN model may use input images of any size thanks to the spatial pyramid pooling method. It keeps track of spatial data in surrounding spatial bins. Both the quantity and size of the bins are defined. Each filter's responses are combined for each spatial bin.

E. Depth-wise and Dilated Separable Convolutions

A depth-wise separable convolution uses a separate kernel for each channel [17]. This is followed by a point-wise convolution using a kernel whose depth is the same as the number of channels. By doing so, the number of parameters and corresponding computational cost are drastically reduced at the price of precision. Dilated convolutions can also be used to expand the receptive field while maintaining relatively low computational costs. This is made possible by applying the kernel to just an input that has gaps caused by pixel skipping. According to the dilation scale, dilated convolutions have the same or fewer parameters than conventional convolutions while providing a larger receptive field.

F. Squeeze and Excitation Networks

In traditional convolution operation, both the spatial and channel-wise information are fused. The focus of [16] is on the correlation between the channel-wise information of the convolution architecture to reinforce the feature representation. To allow the network to perform feature re-calibration, the informative features need to be emphasized as opposed to the less useful ones. After the two blocks of squeeze and excitation, the resultant feature maps will be attentions based on their channel information.

G. Feature Fusion

Combining the features of several frames and scales, many feature fusion strategies were adopted. For instance, [35] introduced a fusion module that performs channel re-ordering across the different frames' corresponding feature maps, followed by a 1x1 convolution and a final concatenation. Before that, concatenation of features or their sum was introduced in [18], where convolutions with kernels of different sizes are used to produce a set of feature maps, followed by concatenation. In [19], the method uses a top-down approach upsampling and addition to combine the pyramid levels. [20] modifies this by concatenating all levels' feature maps to create all the pyramid levels. [21] came later to merge the semantic information on both low-level and high-level, and this is by incorporating the semantic information into the low-level features and the spatial information into the high-level features, then adding the features maps afterwards.

H. Model Ensemble

Deep neural networks are adaptable and scalable to the amount of training data, but with the consequence of being sensitive to their characteristics, resulting potentially in different predictions [31]. One method for reducing the variance of neural network models is to train several models rather than a single model and then aggregate their predictions. This can be done on the same dataset or different datasets within the same domain. Through the object detection literature, many approaches were proposed for this including non-maximum suppression (NMS) [12] and Soft NMS [11].

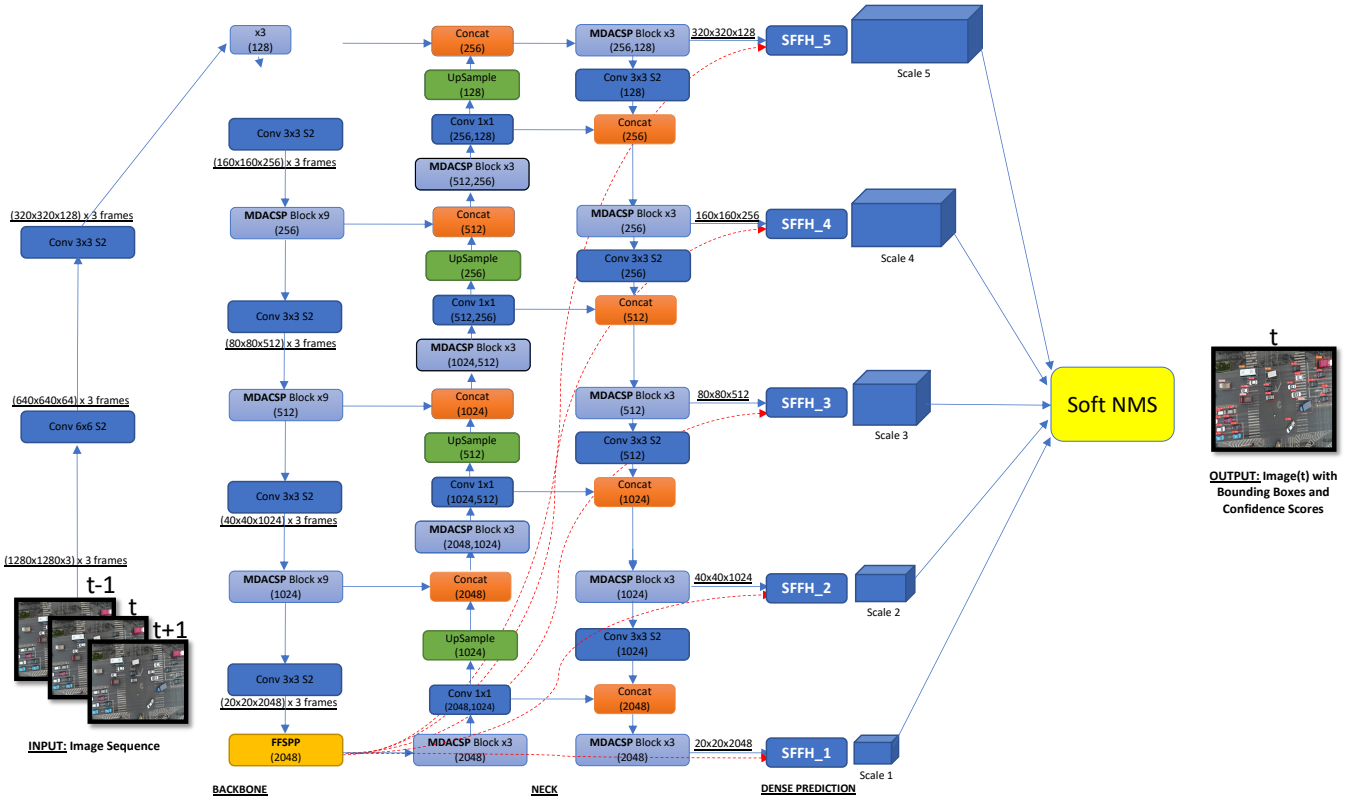

Fig. 2. CEAFFOD Architecture

III. Proposed Approach

A. Architecture

Balancing between memory and computational resources, the single-stage object detection paradigm was adopted [1]. Following the general pipeline of our baseline [30], we introduce different novel components across the backbone, neck and head as shown in Fig. 2. Our contributions are, first, a modified CSPNet block that aims to increase the receptive field, while paying more attention to the significant features. We also introduce a tweaked SPPNet block that aims to have more robust features across scale and time. Thirdly, we introduce a new head that incorporates features from the neck and backbone towards enhancing the dense prediction. The detailed architecture is shown in Fig. 2. Following is more explanation about every contribution.

1) Multi-dilated Attention-based Cross-Stage Partial Connection (MDACSP) Block: MDACSP block is a modification to the Cross-Stage Partial Connection (CSP) [15]. The choice of the CSP block as a base for this modification is for its compatibility with the edge computing applications as with the UAV in terms of the limited computational and memory resources. The modified block utilizes the multi-dilated depth-wise separable convolutions to encode the features of the different scales within the block across the network [17]. This results in better detection of the small and tiny objects. It also applies attention through the channels to produce per-channel modulation weights after aggregating the feature maps across their spatial domain [16]. After that, the features are being fused according to their weights. This results in enhancing the features across the network blocks. The block consists of three stages as shown in Fig. 3.

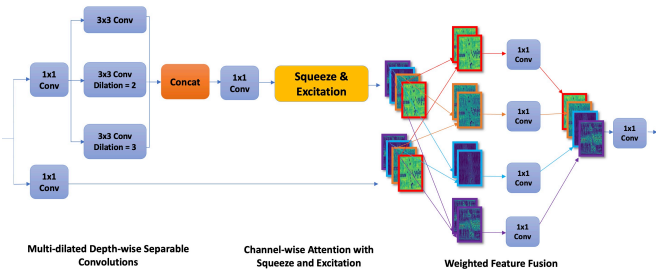

Fig. 3. Mutli-dilated Attention-based Cross-Stage Partial Connection (MDACSP) Block

2) Feature Fusion Spatial Pyramid Pooling (FFSPP) Block: FFSPP block aims to increase the receptive field, by separating the significant features across different scales, and fusing these features all together across the subsequent frames. This results in having features that are more robust to the challenges of the UAV-captured images: scale variation, viewpoint changes, occlusions and others. The block is two stages as shown in Fig. 4.

3) Scaled Feature Fusion Head (SFFH): SFFH is our detection head that fuses the FFSPP enhanced features and the features resulting from the connected MDACSP block specific for this head. Then it is followed by a 1x1 convolution. There are five heads covering different scales from

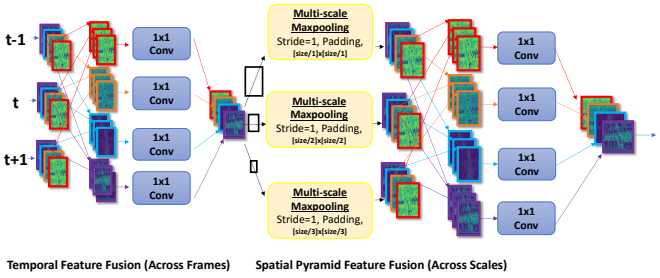

Fig. 4. Feature Fusion Spatial Pyramid Pooling (FFSPP) Block

large to tiny objects. The head design is shown in Fig. 5.

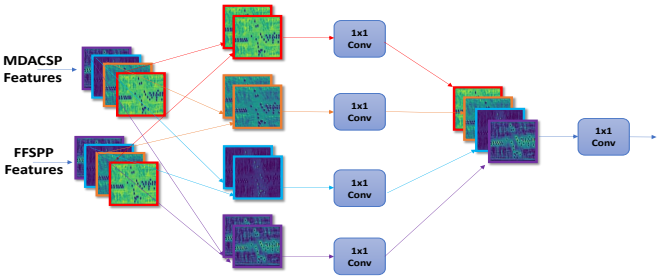

Fig. 5. Scaled Feature Fusion Head (SFFH)

IV. Experiments

A. Datasets and Challenges

To match this level of complexity, three of the top UAV/traffic surveillance datasets have been investigated as in Fig. 1. They covered variety of objects including pedestrians, cars, buses, vans, trucks, bicycles, tricycles, and others. The UAVDT dataset [8] consists of video sequences of 80 thousands frames of complex scenarios and challenges that appear in UAV-based real scenes. This covers variety of high density and small objects, camera motion, weather conditions, different occlusion levels and flying altitudes. To test the robustness of our architecture, we also used the UA-DETRAC benchmark dataset [9]. It includes 100 challenging traffic scenes video sequences taken above highways and intersections. Continuing on the robustness road, we used the VisDrone dataset [10], which includes about 179 thousands frames making 263 video sequences and stationary images of maximum resolutions of 2000x1500 and 3840x2160, respectively. The dataset includes a range of weather and lighting variables to reflect a variety of real-world situations.

B. Evaluation Metrics

For UAV-based vision applications as surveillance and traffic management, mean average precision (mAP) is the main applied object detection metric [48]. An intersection over union (IoU) threshold is used to distinguish between correct and wrong detections during evaluating a model. If the model identified a box IoU larger than this threshold, it is a successful detection. The result used to compute mAP is utilized as the final evaluation index. This measure is the average of all 10 IoU criteria in the range [0.50, 0.95] with a 0.05 constant step size for all categories.

C. Implementation Details

In the UAVDT and UA-DETRAC datasets, we divided the training set into 80% training and 20% validation as proposed by [35]. For the VisDrone dataset, we used the different sets as divided by [10]. Our baseline is the YOLOv4 architecture [30], where we applied our modifications and contributions on the backbone, neck and head as shown in Fig. 2 and Section III. Thus, we partially used pre-trained model [37], few weights could be transferred, on the COCO dataset [38] provided by the baseline. We implemented our work using Pytorch 20.11. Our input images were scaled to 1280x1280, and the data augmentation methods proposed by the best performing model of YOLOv4 were adopted, and they were fixed across the experiments. The used optimizer is Adam [36]. We used an initial learning rate of 0.02. We trained our models for 200 epochs, where the first 10 are used for warm-up. To validate and test the effect of every contribution over the baseline, we first trained the base model end-to-end with every modification individually forming an ablation study, and then they were integrated together for best performance. Seeking maximum performance and robustness, we applied cross ensembling through the three datasets after training all the models. The experiments ran on a NVIDIA RTX A6000 GPU. In addition, to show the edge applicability in the theme of UAVs, the inference accuracy and time are reported on a NVIDIA Jetson Xavier NX board with deploying the models with TensorRT and inferring in FP32 mode with a variation of parameters and input sizes as per the model complexity.

V. Results and Discussion

Table I shows a performance comparison between our method and the current SOTA methods across the datasets of UAVDT, UA-DETRAC and VisDrone. The consistent improvement in all of them shows how efficient and robust our approach is tackling the complexity of the UAV-captured visual data and the variations of the challenges each dataset introduces. Our proposed architecture achieves overall 35.54% mAP more compared to the previous SOTA, FFAVOD [35], in the UAVDT benchmark. For the UA-DETRAC and VisDrone benchmarks, we achieved 93.5% and 42.9% mAP respectively, having an improvement of 5.4% over FFAVOD [35] and 3.47% over DBNet [10]. Fig. 6 and Fig. 7 show the robustness and the accurate object localization of the final model in different complex scenarios. More details are provided in the confusion matrix in Fig. 8, where it is an example showing the detection accuracy for the hard categories that are small scale, dense and suffering from occlusion and viewpoint complexity. The impact of our individual contributions and their integration, across the benchmarks overall and the various object classes, was explored with an ablation study as in tables III, IV and II. The models in these tables are the modifications over the baseline: Model I: MDACSP, Model II: FFSPP, Model III: MDACSP + FFSPP, Model IV: MDACSP + FFSPP + SFFH[3 Heads], Model V: MDACSP + FFSPP + SFFH[4 Heads], Model VI: MDACSP + FFSPP + SFFH[5 Heads], Model VII: MDACSP + FFSPP + SFFH[5 Heads] + Ensemble.

Fig. 6. Qualitative results showing the accurate prediction of the final model and its robustness in different scenarios on NVIDIA Jetson Xavier NX.

TABLE I
COMPARISION BETWEEN CEAFFOD(OURS) AND THE SOTA RESULTS ON UAVDT, UA-DETRAC AND VISDRONE BENCHMARKS

Methods	mAP (%)
UAVDT Benchmark	
LRF-NET [42]	37.81
RetinaNet [43]	38.26
SpotNet [44]	52.80
FFAVOD(previous SOTA) [35]	53.76
CEAFFOD(ours)	89.3 (↑35.54)
UA-DETRAC Benchmark	
JointMODT [45]	83.80
CenterNet [47]	83.48
SpotNet [44]	86.80
FFAVOD(previous SOTA) [35]	88.1
CEAFFOD(ours)	93.5 (↑5.4)
VisDrone Benchmark	
YOLOv3_ReSAM [39]	35.98
ViT-YOLO [40]	38.5
TPH-YOLOv5 [33]	39.18
DBNet (previous SOTA) [29]	39.43
CEAFFOD(ours)	42.9 (↑3.47)

Fig. 7. Qualitative results showing the high accuracy of the final model on an image sequence from the VisDrone benchmark.

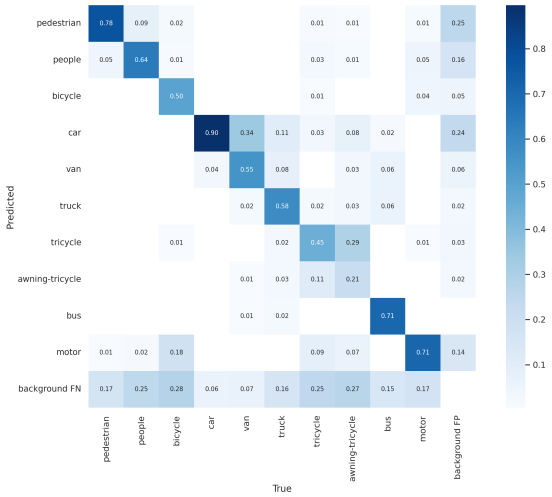

Fig. 8. Confusion matrix made at thresholds of 0.65 and 0.25 for IoU and confidence for the VisDrone benchmark.

The MDACSP block contributed to the results with an improvement of 3.7%, 4.3% and 18.9% mAP over the baseline in the UAVDT, VisDrone and UA-DETRAC benchmarks respectively. The same order of dataset will be implicitly followed in discussing the rest of this section. Investigating the MDACSP block design shown in Fig. 3, the weighted feature fusion across different widened receptive fields and scales worked practically in identifying objects of smaller scales and challenging distortions.

The FFSPP block improved the results by 5.3%, 4.9% and 19.8% mAP over the baseline. With the FFSPP block design as in Fig. 4, the enhanced temporal-spatial feature fusion, across both time and space, yielded in having stronger feature representation for complex scenarios of occlusion and high densities. Integrating both MDACSP and FFSPP blocks yielded better results of 7.4%, 5.9% and 20.6% mAP more than the baseline.

The SFFH block depends on the output of both MDACSP and FFSPP blocks as shown in Fig. 5. Thus, it was tested with both contributions being in the backbone and the neck of the architecture as shown in Fig. 2. Three different experiments were conducted with 3, 4 and 5 SFFH blocks for every dataset. It is noticed that there is a consistent improvement across the three datasets with increasing the number of SFFH blocks from 3 to 5. This is associated with the ability of the block to integrate the MDACSP and FFSPP features by fusing them across the time, space and depth as per each block. For the UAVDT benchmark, the overall improvement caused by the heads themselves adding 3, 4 and 5 heads of the SFFH block is 2.1%, 4.2% and 5% mAP respectively. Similarly, for the VisDrone benchmark, the total mAP increase was 1.2%, 1.7% and 4.3% respectively. As per the same approach with the UA-DETRAC, using 3 heads

TABLE II

COMPARISON OF OUR CEAFFOD CONTRIBUTIONS' PERFORMANCES (MAP%) ON VISDRONE2022 TESTSET-DEV. THIS SERVES AS AN ABLATION STUDY FOR EACH MODIFICATION OVER THE BASELINE.

Model	all	pedestrian	people	bicycle	car	van	trunk	tricycle	awning-tricycle	bus	motor
YOLOv4-baseline	30.9	24.1	17.2	15.7	51	37.7	42.6	22.8	19.7	53.7	24.9
CEAFFOD Model I	35.2	33.5	24.2	19.9	64.2	43	38.6	27.6	15.8	52.3	33.1
CEAFFOD Model II	35.8	33.5	24.5	20.8	64.3	43.4	37.6	29	16.3	55	33.4
CEAFFOD Model III	36.8	36.2	25.5	21.6	65.1	43.6	40	29.8	17.2	54.7	34.5
CEAFFOD Model IV	38	36.8	26.1	21.4	65.4	45.8	41.8	31.5	18.3	57.2	35.5
CEAFFOD Model V	38.5	36.9	27	22.5	66	47.1	41.3	31.1	20.4	55.1	37.2
CEAFFOD Model VI	41.1	40.9	28.7	25.9	69.8	47.9	43.8	33.4	22	59.9	38.4
CEAFFOD Model VII	42.9 (↑ 12)	43.2 (↑ 19)	34.2 (↑ 17)	26.9 (↑ 11.2)	72 (↑ 21)	49.1 (↑ 11.4)	45.3 (↑ 2.7)	33.4 (↑ 10.6)	22.7 (↑ 3)	60.3 (↑ 6.6)	41.8 (↑ 16.9)

TABLE III

COMPARISON OF OUR CEAFFOD CONTRIBUTIONS' PERFORMANCES (MAP%) ON UA-DETRAC TEST-SET AS PER [35]. THIS SERVES AS AN ABLATION STUDY FOR EACH MODIFICATION OVER THE BASELINE.

Model	all	others	car	van	bus
YOLOv4-baseline	67.7	49.25	76.57	59.67	85.11
CEAFFOD Model I	86.6	84.2	82.9	86.1	93
CEAFFOD Model II	87.5	86.5	84.6	87	92
CEAFFOD Model III	88.3	87.6	85.1	87.6	93
CEAFFOD Model IV	90.3	90	87.2	89.3	94.8
CEAFFOD Model V	90.7	90.6	87.6	89.8	95
CEAFFOD Model VI	92.5	92.2	90.1	91.1	96.7
CEAFFOD Model VII	93.5 (↑ 25.8)	93.2 (↑ 43.95)	90.8 (↑ 14.23)	92.5 (↑ 32.83)	97.4 (↑ 12.29)

TABLE IV

COMPARISON OF OUR CEAFFOD CONTRIBUTIONS' PERFORMANCES (MAP%) ON UAVDT TEST-SET AS PER [35]. THIS SERVES AS AN ABLATION STUDY FOR EACH MODIFICATION OVER THE BASELINE.

Model	all	car	truck	bus
YOLOv4-baseline	75.3	69.1	76.2	80.5
CEAFFOD Model I	79	73.6	79.8	83.6
CEAFFOD Model II	80.6	76	80.8	85.1
CEAFFOD Model III	82.7	79	82.5	86.6
CEAFFOD Model IV	84.8	81.4	84.7	88.2
CEAFFOD Model V	86.9	83.3	87.7	89.9
CEAFFOD Model VI	87.7	85.4	87	90.6
CEAFFOD Model VII	89.3 (↑ 14)	87.3 (↑ 18.2)	89.1 (↑ 12.9)	91.4 (↑ 10.9)

gave 2% mAP increase, and 2.4% and 4.2% mAP incease was achieved for adding 4 and 5 heads respectively. Across the benchmarks, we can see that the small and dense classes, such as car and people, were better detected with the adding 5 heads compared to 3 and 4.

Finally, cross-ensemble on top of all the integrated contributions yielded the best results due to the increase of robustness and decreasing the model variance. Final results outperformed the existing SOTA results in the three datasets. Moreover, the results showed the effect of the overall architecture with the contributed modifications over the baseline, archiving an increase over it by 14%, 12% and 25.8% mAP overall across the benchmarks of UAVDT, VisDrone and UA-DETRAC respectively.

CEAFFOD Edge Applicability on NVIDIA Jetson Xavier NX: Table V shows the balanced performance in terms of inference time and accuracy across a variation of different input sizes and parameters as per the model complexity on the UAVDT benchmark. The CEAFFOD lowest-performing edge model is still higher than the previous non-edge SOTA model, and this is with a reasonable inference time of 14.6 ms, equivalent to 68 frames per second (FPS). This comes with a memory cost of 19.2 million parameters. The accuracy of the CEAFFOD models increases with the model complexity reaching 81.2% mAP processing 16 FPS with 36.7 million parameters. The results show a desirable variety of space-time complexity combinations suiting the different edge-oriented design and deployment requirements.

TABLE V

THE INFERENCE PERFORMANCE OF THE CEAFFOD MODELS ON THE UAVDT BENCHMARK ON THE NVIDIA JETSON XAVIER NX BOARD. THE PERFORMANCE IS MEASURED WITH ACCURACY (MAP (%)) AND TIME (MS) WITH A VARIATION OF PARAMETERS AND INPUT SIZES ACCORDING TO THE MODEL COMPLEXITY OVER THE BASELINE.

Model	Input Size	#Params (M)	Time (ms)	mAP (%)
CEAFFOD Model I	640^2	19.2	14.7	61.9
CEAFFOD Model II	640^2	21.4	19.3	56.8
CEAFFOD Model III	640^2	26.0	20.1	68.7
CEAFFOD Model IV	640^2	27.4	26.9	63.6
CEAFFOD Model V	1280^2	29.2	54.1	71.4
CEAFFOD Model VI	1280^2	30.6	59.0	67.0
CEAFFOD Model VII	1280^2	36.7	61.6	81.2

VI. CONCLUSION

In this paper, we propose three novel blocks modifying the single-stage YOLOv4 object detection baseline to tackle the challenges of the UAV perception nature. Our work yielded the best-performing results across three of the top UAV/traffic surveillance datasets: UAVDT, UA-DETRAC and VisDrone. We have carried out an ablation study with our three proposed blocks: MDACSP, FFSPP and SFFH and its variations, and they all increased the overall performance consistently. Cross-ensemble enhanced the results after integrating all the contributions and training the models in an end-to-end manner on the three datasets. Testing the CEAFFOD models on NVIDIA Jetson Xavier NX board shows SOTA-outperforming accuracy with an applicable balance of memory and speed. On top of these promising quantitative results, our qualitative results show how our approach is more robust and efficient in the UAV-captured image sequences in complex scenarios.

VII. ACKNOWLEDGMENT

This work is part of the collaborative project "Intelligent object detection, dynamic scene and activity recognition for real-time UAV applications" between Technology Innovation Institute (TII) and Mohamed bin Zayed University of Artificial Intelligence (MBZUAI).

REFERENCES

[1] Ramachandran, A. & Sangaiah, A. A review on object detection in unmanned aerial vehicle surveillance. *International Journal Of Cognitive Computing In Engineering (2021)*.

[2] Jain, A., Ramaprasad, R., Narang, P., Mandal, M., Chamola, V., Yu, F. & Guizan, M. AI-Enabled Object Detection in UAVs: Challenges, Design Choices, and Research Directions. *IEEE Network*. **35**, 129-135 (2021)

[3] Blaschke, T., Lang, S. & Hay, G. Object-Based Image Analysis: Spatial Concepts for Knowledge-Driven Remote Sensing Applications. *Geographic Object-Based Image Analysis (GEOBIA)*.. (2008,1)

[4] Wu, X., Li, W., Hong, D., Tao, R. & Du, Q. Deep Learning for Unmanned Aerial Vehicle-Based Object Detection and Tracking: A survey. *IEEE Geoscience And Remote Sensing Magazine*. **10**, 91-124 (2022)

[5] Sun, W., Dai, L., Zhang, X., Chang, P. & He, X. RSOD: Real-time small object detection algorithm in UAV-based traffic monitoring. *Applied Intelligence*. **52**, 8448-8463 (2022,6,1), https://doi.org/10.1007/s10489-021-02893-3

[6] Cazzato, D., Cimarelli, C., Sanchez-Lopez, J., Voos, H. & Leo, M. A Survey of Computer Vision Methods for 2D Object Detection from Unmanned Aerial Vehicles. *Journal Of Imaging*. **6**, 78 (2020,8), https://www.mdpi.com/2313-433X/6/8/78

[7] Osco, L., Marcato Junior, J., Marques Ramos, A., Castro Jorge, L., Fatholahi, S., Andrade Silva, J., Matsubara, E., Pistori, H., Gonçalves, W. & Li, J. A review on deep learning in UAV remote sensing. *International Journal Of Applied Earth Observation And Geoinformation*. **102** pp. 102456 (2021,10), https://linkinghub.elsevier.com/retrieve/pii/S030324342100163X

[8] Yu, H., Li, G., Zhang, W., Huang, Q., Du, D., Tian, Q. & Sebe, N. The Unmanned Aerial Vehicle Benchmark: Object Detection, Tracking and Baseline. *International Journal Of Computer Vision*. **128**, 1141-1159 (2020,5), http://link.springer.com/10.1007/s11263-019-01266-1

[9] Wen, L., Du, D., Cai, Z., Lei, Z., Chang, M., Qi, H., Lim, J., Yang, M. & Lyu, S. UA-DETRAC: A new benchmark and protocol for multi-object detection and tracking. *Computer Vision And Image Understanding*. **193** pp. 102907 (2020,4), https://linkinghub.elsevier.com/retrieve/pii/S1077314220300035

[10] Du, Y., Wan, J., Zhao, Y., Zhang, B., Tong, Z. & Dong, J. GIAOTracker: A comprehensive framework for MCMOT with global information and optimizing strategies in VisDrone 2021. *2021 IEEE/CVF International Conference On Computer Vision Workshops (ICCVW)*. pp. 2809-2819 (2021,10), https://ieeexplore.ieee.org/document/9607520/

[11] Bodla, N., Singh, B., Chellappa, R. & Davis, L. Soft-NMS — Improving Object Detection with One Line of Code. *2017 IEEE International Conference On Computer Vision (ICCV)*. pp. 5562-5570 (2017,10), http://ieeexplore.ieee.org/document/8237855/

[12] Hosang, J., Benenson, R. & Schiele, B. Learning Non-maximum Suppression. *2017 IEEE Conference On Computer Vision And Pattern Recognition (CVPR)*. pp. 6469-6477 (2017,7), http://ieeexplore.ieee.org/document/8100168/

[13] Nousi, P., Mademlis, I., Karakostas, I., Tefas, A. & Pitas, I. Embedded UAV Real-Time Visual Object Detection and Tracking. *2019 IEEE International Conference On Real-time Computing And Robotics (RCAR)*. pp. 708-713 (2019)

[14] Ravi, N. & El-Sharkawy, M. Real-Time Embedded Implementation of Improved Object Detector for Resource-Constrained Devices. *Journal Of Low Power Electronics And Applications*. **12**, 21 (2022,4), https://www.mdpi.com/2079-9268/12/2/21

[15] Wang, C., Mark Liao, H., Wu, Y., Chen, P., Hsieh, J. & Yeh, I. CSPNet: A New Backbone that can Enhance Learning Capability of CNN. *2020 IEEE/CVF Conference On Computer Vision And Pattern Recognition Workshops (CVPRW)*. pp. 1571-1580 (2020,6), https://ieeexplore.ieee.org/document/9150780/

[16] Hu, J., Shen, L., Albanie, S., Sun, G. & Wu, E. Squeeze-and-Excitation Networks. *IEEE Transactions On Pattern Analysis And Machine Intelligence*. **42**, 2011-2023 (2020,8), https://ieeexplore.ieee.org/document/8701503/

[17] Kim, S. & Lee, H. Lightweight Stacked Hourglass Network for Human Pose Estimation. *Applied Sciences* 6497 (2020)

[18] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. & Rabinovich, A. Going deeper with convolutions. *Proceedings Of The IEEE Conference On Computer Vision And Pattern Recognition*. pp. 1-9 (2015)

[19] Lin, T., Dollár, P., Girshick, R., He, K., Hariharan, B. & Belongie, S. Feature pyramid networks for object detection. *Proceedings Of The IEEE Conference On Computer Vision And Pattern Recognition*. pp. 2117-2125 (2017)

[20] Li, Z. & Zhou, F. FSSD: feature fusion single shot multibox detector. *ArXiv Preprint ArXiv:1712.00960*. (2017)

[21] Zhang, Z., Zhang, X., Peng, C., Xue, X. & Sun, J. Exfuse: Enhancing feature fusion for semantic segmentation. *Proceedings Of The European Conference On Computer Vision (ECCV)*. pp. 269-284 (2018)

[22] Yu, J., Jiang, Y., Wang, Z., Cao, Z. & Huang, T. UnitBox: An Advanced Object Detection Network. *Proceedings Of The 24th ACM International Conference On Multimedia*. pp. 516-520 (2016), https://doi.org/10.1145/2964284.2967274

[23] Girshick, R., Donahue, J., Darrell, T. & Malik, J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. *2014 IEEE Conference On Computer Vision And Pattern Recognition*. pp. 580-587 (2014,6), http://ieeexplore.ieee.org/document/6909475/

[24] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. & Berg, A. SSD: Single Shot MultiBox Detector. *Computer Vision – ECCV 2016*. pp. 21-37 (2016)

[25] Mittal, P., Singh, R. & Sharma, A. Deep learning-based object detection in low-altitude UAV datasets: a survey. *Image And Vision Computing*. (2020,12)

[26] Redmon, J., Divvala, S., Girshick, R. & Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. *2016 IEEE Conference On Computer Vision And Pattern Recognition (CVPR)*. pp. 779-788 (2016,6), http://ieeexplore.ieee.org/document/7780460/

[27] Howard, A., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M. & Adam, H. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. (2017,4)

[28] He, K., Zhang, X., Ren, S. & Sun, J. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. *IEEE Transactions On Pattern Analysis And Machine Intelligence*. **37**, 1904-1916 (2015,9), http://ieeexplore.ieee.org/document/7005506/

[29] Cao, Y., He, Z., Wang, L., Wang, W., Yuan, Y., Zhang, D., Zhang, J., Zhu, P., Van Gool, L., Han, J., Hoi, S., Hu, Q. & Liu, M. VisDrone-DET2021: The Vision Meets Drone Object Detection Challenge Results. *Proceedings Of The IEEE/CVF International Conference On Computer Vision (ICCV) Workshops*. pp. 2847-2854 (2021,10)

[30] Bochkovskiy, A., Wang, C. & Liao, H. YOLOv4: Optimal Speed and Accuracy of Object Detection (2020)

[31] Ganaie, M., Hu, M., Malik, A., Tanveer, M. & Suganthan, P. Ensemble deep learning: A review. *Engineering Applications Of Artificial Intelligence*. **115** pp. 105151 (2022,10), https://linkinghub.elsevier.com/retrieve/pii/S095219762200269X

[32] Liu, M., Wang, X., Zhou, A., Fu, X., Ma, Y. & Piao, C. UAV-YOLO: Small Object Detection on Unmanned Aerial Vehicle Perspective. *Sensors*. **20** pp. 2238 (2020,4)

[33] Zhu, X., Lyu, S., Wang, X. & Zhao, Q. TPH-YOLOv5: Improved YOLOv5 Based on Transformer Prediction Head for Object Detection on Drone-Captured Scenarios. *Proceedings Of The IEEE/CVF International Conference On Computer Vision (ICCV) Workshops*. pp. 2778-2788 (2021,10)

[34] Jung, H. & Choi, G. Improved YOLOv5: Efficient Object Detection Using Drone Images under Various Conditions. *Applied Sciences*. **12** (2022), https://www.mdpi.com/2076-3417/12/14/7255

[35] Perreault, H., Bilodeau, G., Saunier, N. & Héritier, M. FFAVOD: Feature fusion architecture for video object detection. *Pattern Recognition Letters*. **151** pp. 294-301 (2021,11), https://linkinghub.elsevier.com/retrieve/pii/S016786552100307X

[36] Kingma, D. & Ba, J. Adam: A Method for Stochastic Optimization. *International Conference On Learning Representations*. (2014,12)

[37] Huang, S. & Liu, Q. Addressing scale imbalance for small object detection with dense detector. *Neurocomputing*. **473** pp. 68-78 (2022,2), https://linkinghub.elsevier.com/retrieve/pii/S0925231221018191

[38] Lin, T., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P. & Zitnick, C. Microsoft COCO: Common Objects in Context. *Computer Vision – ECCV 2014*. pp. 740-755 (2014)

[39] Liu, B., Luo, H., Wang, H. & Wang, S. YOLOv3_ReSAM: A Small-Target Detection Method. *Electronics*. **11** (2022), https://www.mdpi.com/2079-9292/11/10/1635

[40] Zhang, Z., Lu, X., Cao, G., Yang, Y., Jiao, L. & Liu, F. ViT-YOLO:Transformer-Based YOLO for Object Detection. *2021 IEEE/CVF International Conference On Computer Vision Workshops (ICCVW)*. pp. 2799-2808 (2021)

[41] Xu, C., Hong, X., Yao, Y., Shen, H., Ma, Q. & Jiang, H. Multi-Scale Region-based Fully Convolutional Networks. *2020 IEEE International Conference On Power, Intelligent Computing And Systems (ICPICS)*. pp. 500-505 (2020,7), https://ieeexplore.ieee.org/document/9202049/

[42] Wang, T., Anwer, R., Cholakkal, H., Khan, F., Pang, Y. & Shao, L. Learning Rich Features at High-Speed for Single-Shot Object Detection. *2019 IEEE/CVF International Conference On Computer Vision (ICCV)*. pp. 1971-1980 (2019,10), https://ieeexplore.ieee.org/document/9008401/

[43] Lin, T., Goyal, P., Girshick, R., He, K. & Dollar, P. Focal Loss for Dense Object Detection. *IEEE Transactions On Pattern Analysis And Machine Intelligence*. **42**, 318-327 (2020,2), https://ieeexplore.ieee.org/document/8417976/

[44] Perreault, H., Bilodeau, G., Saunier, N. & Heritier, M. SpotNet: Self-Attention Multi-Task Network for Object Detection. *2020 17th Conference On Computer And Robot Vision (CRV)*. pp. 230-237 (2020,5), https://ieeexplore.ieee.org/document/9108685/

[45] Huang, K. & Hao, Q. Joint Multi-Object Detection and Tracking with Camera-LiDAR Fusion for Autonomous Driving. *2021 IEEE/RSJ International Conference On Intelligent Robots And Systems (IROS)*. pp. 6983-6989 (2021,9), https://ieeexplore.ieee.org/document/9636311/

[46] Fu, Z., Chen, Y., Yong, H., Jiang, R., Zhang, L. & Hua, X. Foreground Gating and Background Refining Network for Surveillance Object Detection. *IEEE Transactions On Image Processing*. **28**, 6077-6090 (2019)

[47] Duan, K., Bai, S., Xie, L., Qi, H., Huang, Q. & Tian, Q. CenterNet: Keypoint Triplets for Object Detection. *2019 IEEE/CVF International Conference On Computer Vision (ICCV)*. pp. 6568-6577 (2019,10), https://ieeexplore.ieee.org/document/9010985/

[48] Zaidi, S., Ansari, M., Aslam, A., Kanwal, N., Asghar, M. & Lee, B. A survey of modern deep learning based object detection models. *Digital Signal Processing*. **126** pp. 103514 (2022,6), https://linkinghub.elsevier.com/retrieve/pii/S1051200422001312