2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

Learning Multi-Scale Context Mask-RCNN Network for Slant Angled Aerial Imagery in Instance Segmentation in a Sim2Real setup

Qiranul Saadiyean[1], S P Samprithi[2] and Suresh Sundaram[3]

Abstract*— While instance segmentation models excel at object detection in satellite imagery, their performance drops when applied to slant-angled aerial images due to occlusion and scale variation. This is mainly caused by a lack of training data for such diverse viewpoints and scales. To address this limitation, we propose the Sim2Real-based Multi-Scale Context Mask-RCNN (MSC-RCNN) network, specifically designed for slant-angled aerial imagery. Sim2Real-based transfer learning is adapted to compensate for the limited availability of real-world slant-angle training data. A synthetic dataset is generated using Unreal Engine, detailing the methodology of replicating the real-world scene, for producing diverse slant-angle drone datasets with various weather conditions and backgrounds. The model leverages two distinct feature pyramid backbones, with one incorporating dilated convolutions to address large-scale objects and the other optimized for regular convolutions. Their outputs are fused to effectively detect objects across various scales and angles. Through experiments, it was demonstrated that incorporating this synthetic data significantly reduces reliance on real data while maintaining high mean Average Precision (mAP) scores. Compared to the baseline Mask R-CNN, the proposed approach with Sim2Real adaptation and the MSC-RCNN architecture achieves a remarkable 7.6% performance improvement in instance segmentation accuracy with only a 6% increase in model size. Code can be found at: *https://github.com/MSC-RCNN

I. INTRODUCTION

Recent advancements in aerial vehicles (UAVs) have led to their widespread adoption in various applications, including surveillance [40], fire detection [18], [7], [1], traffic planning [19], infrastructural navigation, disaster management [41], and perimeter defense [34]. In disaster management and surveillance, surveying an area from an angle ahead rather than directly from overhead, can significantly increase the coverage area, enabling more efficient decision-making and planning. However, this forward-looking approach, as seen in Fig. 1, results in slant-angled oblique images. While conventional methods involve orthorectification of such images for image processing, they become ineffective for long-range aerial imagery due to large angles involved [30]. This leads to issues like scale changes and occlusions, causing significant distortion and information loss during orthorectification. Hence, there is a critical need to develop new deep learning models specifically designed to handle these challenging high-resolution aerial slant angled images. Furthermore, most existing object detection algorithms for computer vision tasks [42], primarily Convolutional Neural Networks (CNNs), require vast amounts of annotated data. However, creating such datasets is often a manual, time-consuming, and expensive process. These real-world datasets are often limited in availability and may under represent certain classes, hindering the model's generalizability. To address these data limitations, synthetic data generation using simulations has emerged as a promising alternative [31][29]. These synthetic datasets offer the advantage of automatic annotation and data generation, saving time and resources.

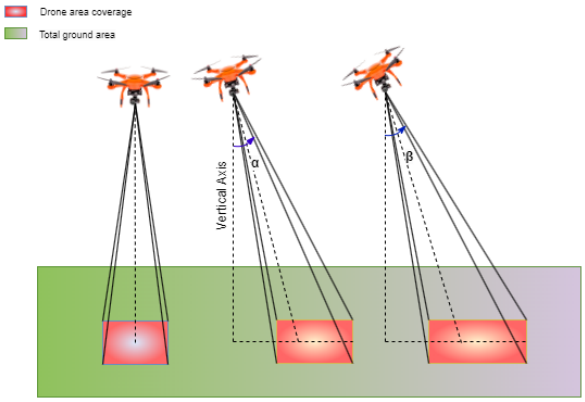

Fig. 1: Drone imagery representation. Here 15< α <45 represents small slant angle , 50< β <75 higher slant angle images. As seen from the image, increasing the slant angle increases the coverage area.

However, solely relying on synthetic data introduces a domain gap, meaning the model performs well on simulated data but struggles with real-world scenarios [33]. To bridge this gap, paired real datasets with shared features are often used as a baseline for performance evaluation and domain gap minimization. However, acquiring real-world data for high-slant angles presents its own set of challenges, including: Occlusion and scale variations, Limited availability and under representation: Real-world datasets are often limited in size and may not adequately represent the diversity of real-world scenarios. Additionally, certain classes within the data might be underrepresented, leading to biases in the model's performance.

To address these limitations, Sim2Real[14] based transfer learning approach is adopted for slant-angled aerial imagery. In Sim2Real approach, a model is first trained on a synthetic dataset. This not only saves time and resources compared to

*This work was funded by MeitY (Ministry of Electronics and Information Technology).

[1]Qiranul Saadiyean is with Department of Aerospace Engineering, Indian Institute of Science, qiranuls@iisc.ac.in

[2]S P Samprithi is with Department of Electronics and Communication Engineering, PES University, spsamprithi@pes.pesu.edu

[3]Suresh Sundaram is with the Department of Aerospace Engineering, Indian Institute of Science, vvsuresh@iisc.ac.in

manual real-world data annotation, but also solves the issue of data scarcity. Subsequently, the model is fine-tuned on a limited real-world dataset to bridge the domain gap and improve its performance on real-world tasks. This approach leverages the efficiency and scalability of synthetic data generation while ensuring the model generalizes well to real-world scenarios.

Further the mainstream object detection algorithms are divided into two classes: single-stage detectors and two-stage detectors. Single-stage detectors perform detection in a single stage, directly proposing class probability and box-level predictions of the object [27], [24]. On the other hand, two-stage detectors, such as Fast R-CNN [10], Faster R-CNN [28], Mask R-CNN [12], follow a more intricate approach. They first generate candidate regions or bounding boxes likely to contain objects of interest in a region proposal generation stage, followed by object classification and refinement within these proposed regions.

While in traditional object detection tasks, bounding boxes are used to localize objects, they lack the capability to differentiate between multiple instances of the same class that might overlap or be occluded. To address this limitation, instance segmentation [12], [11] is adopted due to its ability to provide granular information about object boundaries, size and individual instances within a scene. This detailed segmentation enables better understanding and analysis of aerial imagery, allowing for more reliable object detection, tracking, and subsequent decision-making in various domains, including surveillance and monitoring. Various datasets have been created in recent years to focus specifically on aerial object detection from drones. Most of these existing datasets in aerial imagery are the overhead imagery which consists of images taken from top view. One of the primarily used datasets for aerial surveillance involves the DOTA dataset [37]. It contains multiple object classes typically found in aerial imagery, including airplanes, ships, vehicles, buildings, bridges, etc. Since the dataset has been taken from satellite view and thus fails to address images of varying scale, the object detection performance deteriorates at slant angles, when trained on such datasets. While orthorectification, a common technique for correcting geometric distortions, is feasible for datasets like DOTA due to their small viewing angles, it becomes impractical for higher slant angle aerial imagery due to thermal variations and occlusion. Thus in view of all these limitations, we introduce a Multi scale context aware Mask R-CNN (MSC-RCNN), a deep learning framework designed to tackle the challenges associated with slant-angled aerial imagery in instance segmentation in Sim2Real setup. MSC-RCNN employs a dual Feature Pyramid Network (FPN) [21] architecture, where one FPN incorporates dilated convolutions [39] while the other does not. These parallel FPNs are combined with a shared Regional Proposals Network (RPN) and Mask R-CNN head. The incorporation of dilated convolutions enhances the model's capability to detect objects of various scales and sizes without compromising resolution or computational efficiency. It has been trained on synthetic dataset and fine-tuned on real data to bridge the domain gap, ensuring robust performance in real-world scenarios. By introducing MSC-RCNN, we aim to push the boundaries of aerial object detection and instance segmentation, particularly in the face of the unique challenges posed by slant-angle aerial imagery. This advancement paves the way for more accurate and reliable analyses in surveillance and disaster management applications. The main contributions of the proposed work are :

1. A novel simulated aircraft dataset generation approach for the specific challenge of slant angled aerial imagery inclusive of various weather conditions.

2. A Multi-Scale context aware Mask R-CNN is proposed for the task of instance segmentation of slant angled aerial imagery.

3. Extensive experiments to validate the proposed model against simulated and real slant angled aerial imagery.

II. Related Works

A. Aerial object detection datasets

Various datasets have been released based on aerial object detection from drones. These datasets can be categorized into two main types based on their primary goals: (i) Object detection (ii) Object segmentation.

DOTA [37], iSAID [35], AID [38], HRSC2016 [25], and LandCover [2] are mainly composed of satellite images with high resolution. CARPK [15], contains images of parking spaces using a drone in different locations, the Aerial Image Dataset (AID) facilitates aerial scene classification tasks. High resolution ship collections 2016 (HRSC2016) [25] dataset contains images from two scenarios including ships on sea and ships in-shore (70 sea images with 90 samples and 991 sea-land images with 2886 samples) collected from six famous harbours. Commonly utilized for semantic segmentation tasks, the LandCover.ai (Land Cover from Aerial Imagery) [2] dataset is widely employed for automatic mapping of buildings, woodlands, water, and roads from aerial images. The Car Parking Lot Dataset (CARPK) [15] comprises approximately 90,000 car instances gathered from four distinct parking lots using drones. The dataset is specifically focused on accurately counting the number of cars within a designated region, such as a parking lot. However, this does not account for any oblique angle images which is a critical problem for counting the cars or objects for surveillance. The UAVDT [8] and VisDrone [4] datasets were acquired using drones and are widely used for object detection in traffic-related applications. While they excel in handling occlusion and small-scale object detection, they do not provide support for object segmentation, which is crucial for estimating object size and other derived parameters. It is unsuitable for segmenting objects in special scenarios such as driving and surveillance due to the viewing angle, small objects, and occlusion. All of the mentioned dataset fail to address slant angle scenarios and scale variation.

B. Algorithms

Ross Girshick et al. [10] proposed a fast region-based convolutional network (Fast R-CNN) for object detection.

Shaoqing Ren et al. [28] advances Fast R-CNN by introducing RPN (Region Proposal Network) which shares full-image convolutional features in the detection network hence enabling region proposals for object detection. Kaiming He et al. proposed methods [13] for image recognition and [12] for instance segmentation, bounding box detection, and person key-point detection. Mask R-CNN extends Faster R-CNN by adding a branch for predicting an object mask along with the existing branch for bounding box recognition. Mask R-CNN efficiently detects objects in an image by generating the segmentation masks simultaneously.

C. Sim2Real approaches

Horváth et al. [14] proposed a Sim2Real transfer learning method using domain randomization for object detection in robotics, effectively bridging the reality gap between synthetic and real data. Jacob Shermeyer et al. [31] introduced the RarePlanes dataset, which focuses on using synthetic data to generate object detection dataset for satellite imagery. Building upon this concept, Xiaomin Lin et al. [23] proposed the SeaDroneSim dataset, utilizing 3D models of objects to generate synthetic data. Most of the work mentioned involves Sim2Real for industrial setup and for top down satellite views. However, none of the works have tried to address the issue of instance segmentation for slant angle images. Hence to address these issues Sim2Real approach is used to generate slant angled aerial imagery dataset for segmentation task.

III. SlantSim Dataset

A specialized dataset comprising both real and synthetic images captured from a drone at an slant angle was created. The drone, as illustrated in Fig. 2b, was operated at altitudes ranging from 15 to 25 meters, and images were acquired from this slanted perspective. In case of the synthetic dataset, as shown in Fig. 2a, parameters were configured to mirror real-world conditions, ensuring an accurate representation of the physical environment. The ***dataset***[1] is composed of four distinct classes: fighter jets, helicopters, carrier planes, and passenger planes. In the real domain, the distribution of instances across these classes is as follows: fighter jets (1196 instances), helicopters (483 instances), carrier planes (5 instances), and passenger planes (65 instances). Notably, the number of instances is relatively low, and they exhibit a long-tail distribution among the classes.

The simulated dataset was expanded to include 854 instances of fighter jets, 385 instances of helicopters, 212 instances of carrier planes, and 548 instances of passenger planes. This dataset was thoughtfully crafted to augment instances, particularly for classes that had limited representation in the existing real domain. As shown in Fig. 3, the generated dataset is noteworthy for its ability to encompass scenarios, such as rain, storm, sunset, sunrise etc. which is challenging to capture in real dataset.

[1]The dataset can be found at: SlantSim-Synthetic-Dataset

(a) Simulated scene with drone. (b) Real drone.

Fig. 2: Drone visualization. The green box in (b) represents the drone at slant angle.

A. Model Generation

Open source 3D models have be downloaded from internet and used for simulation. These models include 3D cad models and photogrammetric [26] models with their textures. The models used are of real world scale for accurate scene representation. The models once imported to unreal engine, can be placed at any location in the map. The classes were selected to match our training dataset for accurate comparison.

B. Simulation

In our Sim2Real approach, Ue4 (Unreal Engine 4) is utilized to simulate the environment. It is a widely used and powerful game engine developed by Epic Games [9]. Camera parameters can be modeled according to the real camera specifications. To configure camera parameters, such as noise and other settings, we have made use of of built-in C++ libraries provided by Unreal Engine. This allows precise control and customization of camera setup for accurate simulation. The sensor parameter which are varied include: Resolution, F-stop, Shutter Speed, Horizontal FOV (Field of view), ISO, Gamma etc.

One of the key advantages of using Unreal Engine is the ability to generate both semantic and instance segmentation mask. By leveraging the engine's capabilities, one can create datasets that contains various challenging conditions that are difficult to replicate in real-world data collection. For instance as shown in Fig. 3, simulating low-light conditions, heavy rainfall, or lens blur caused by rain, which are often impractical or challenging to capture in real-world scenarios

C. Scene Generation

Maps can be generated using the unreal engine for accurate simulation and comparison. The map generated in our simulation is similar to the test dataset for proper evaluation. The map can be manipulated to include various obstructions. Various environmental conditions are simulated including daylight, nighttime, rain, storms, and lens occlusion, among others. This diversity in conditions enables us to thoroughly evaluate the robustness and generalization capabilities of our object detection algorithms. Once scene is simulated, dataset is generated, with various weather conditions, scale and classes configuration. The background is also varied for better generalization.

Fig. 3: Synthetic dataset generation using Ue4, with varied environmental conditions and angles.

By utilizing the Unreal Engine and simulating a wide range of environmental conditions, we aim to generate high-quality datasets that closely resemble real-world scenarios. These datasets will serve as valuable resources for training and testing object detection models, ultimately enhancing their performance in real-world applications.

IV. MSC-RCNN

In this section Multi-Scale Context Mask R-CNN (MSC-RCNN), a deep learning framework to detect instances present in slant angle aerial images is presented. The Fig. 4 represents the proposed architecture.

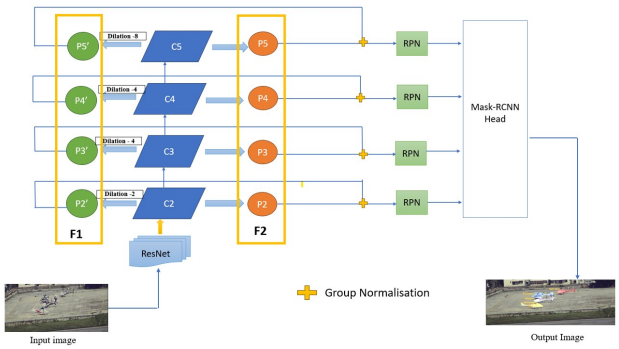

Fig. 4: Architecture of the proposed network - MSC-RCNN.

A. Overview of MSC-RCNN

The proposed MSC-RCNN architecture is an extension of the existing Mask R-CNN architecture as shown in Fig. 4. The backbone of MSC-RCNN consists of two feature pyramid networks instead of the conventional single FPN (Feature Pyramid Network) backbone. The two FPN's share the same ResNet backbone. The FPN's differs with respect to with and without the presence of dilated convolutions in the last layers of the residual blocks of Resnet C2, C3, C4, C5 that are used to construct the FPN. The output from every level of both the FPN backbones are concatenated with each other, and then passed to the common RPN and Mask R-CNN head. The criteria used for concatenating the two FPN's is explained in detail in section 4.4.

B. F1 FPN

The instances in context are of multiple scales, hence the nature of the dilated convolutions to support expansion of the receptive fields without the loss of resolution and increase in computation cost is utilised. It is necessary to maintain both resolution and context in order to accurately detect instances. The dilated convolutions of varying rate was introduced in the last layers of the ResNet residual blocks C2, C3, C4, C5 which would generate FPN levels P2, P3, P4, P5 respectively. The dilated convolutions was introduced in all the residual blocks to capture both local and global feature information to accurately detect instances. The dilation rate across different layers corresponding to different levels of the FPN backbone varied as 2, 4, 4, 8 corresponding to the FPN levels P2, P3, P4, P5. The dilation rate was increased with increase in the depth of the network, in-order to facilitate the model to effectively capture the high level semantic features along with multi-scale contextual information. The FPN's already facilitate the detection of multi-scale objects, but the further addition of dilated convolution at different levels of FPN, boosts the performance of the model. Hence, it was observed that the introduction of dilated convolutions boosted the performance of detection of small objects in low slant angle images whereas it helped in the detection of multi-scale objects in case of high slant angle images.

C. F2 FPN

The layers in F2 FPN are constructed using the conventional convolutional layers in the C2, C3, C4, C5 with fixed receptive field size. The addition of dilated convolutions result in detection of instances of different scales , but also result in false alarms. The normal convolutions were better of than dilated convolutions at picking instances in some cases such as low slant images. Hence, in order to effectively counter the features convolved by both F1 and F2 FPN, the output from both the backbones were fused.

D. Fusion of F1 and F2 FPN

The outputs produced by different levels of FPN are P2, P3, P4, P5, they share the same spatial resolution as the ResNet residual blocks from which they are convolved

through lateral connections and addition of upsampled layers of higher level.

The layers from the corresponding levels of the two FPNs are concatenated with each other. That is P2 from F1 is concatenated with P2 of F2 and so on. The concatenated layer is first convolved using 1x1 convolutions before passing it to the Group normalisation (GN) [36] algorithm to increase the interactions between the concatenated layers. GN is a type of normalisation technique, where the input feature channels are divided into groups, along which the mean and variances are calculated. The number of groups G was equated to two, since the feature maps from two different FPN backbones were concatenated.

The adoption of GN layers, ensures that the features are effectively combined, by making sure the negative effects of a feature from one FPN backbone does not supersede the positive/accurate detections from another backbone. The obtained outputs are then passed through 1x1 convolution, which reduces the number of channels to half its original number, to match its feature channels to perform further operations. The effect of GN on model evaluation metrics is discussed in section 6.2.

V. Experiments

In this section, model training implementation details and its performance with respect to other models are compared.

A. Implementation details

The Mask R-CNN model with ResNet-101 backbone is considered as the baseline model of the proposed work. All the backbone considered in the experimentation was pre-trained using ImageNet. The proposed MSC-RCNN was optimised using Adam optimiser with a learning rate of 0.0001. A consistent learning rate was maintained for all the experiments. Additionally, all the layers of the backbone were trained, since it was observed that freezing of some layers did not improve the model performance. All the models were trained until their convergence which was around 40 epochs. The models were trained on NVIDIA Quadro RTX 5000 with 16GB memory.

Params	Model	AP	AP50	AP75	*APs*	*APm*	*APl*
62.8	Mask R-CNN[12]	48.2	87	39.7	33.9	51.4	53.7
79.24	MS R-CNN[16]	49	87.8	41	35.5	51.0	56.7
96.024	Cascade Mask R-CNN[3]	48.7	87.9	60.6	26.1	47.9	57.9
74.91	PointRend[20]	48.1	89.1	57.2	24.6	51.2	54.1
66.617	**MSC-RCNN**	51.4	89.9	65.4	36.5	51.6	58.9

TABLE I: Results of proposed MSC-RCNN model with respect to the state of the art models in real domain.

Params	Model	AP	AP50	AP75	APs	APm	API
62.8	Mask R-CNN[12]	50.8	88.3	64.6	34.8	55.1	58.8
79.24	MS R-CNN[16]	51.6	89.5	65	32.5	50.1	57.9
96.024	Cascade Mask R-CNN[3]	55.4	89	66.4	28.6	53.5	65.7
74.91	PointRend[20]	53.1	88.9	65.6	27.1	52.3	62
66.617	**MSC-RCNN**	55.8	91	64.7	36.9	55.1	63.5

TABLE II: Results of proposed MSC-RCNN model with respect to the state of the art models in Sim2Real Fine-tuning.

The Mask R-CNN was chosen to be the baseline model since the number of parameters of the proposed MSC-RCNN and the latter differs by relatively lesser difference compared to other state of the art models. Also, the proposed MSC-RCNN model is an extension of Mask R-CNN with changes in backbone. This Sim2Real approach with fewer samples of real and dilated convolution in transformer [17] may help in slant angle aerial imagery. But since transformers require large amount of data for training, it has not been addressed in this work. The Fig. 5 represents the qualitative results.

The COCO Evaluation metrics [22] are used to evaluate the models performance in terms of Average Precision (AP). From the Table I and Table II, it can be observed that the proposed MSC-RCNN model outperforms the baseline by 3.2 % and by 5 % when trained on real and Sim2Real FT (Fine tuning) respectively. Further, it is observed that the performance of both the baseline and the proposed MSC-RCNN model increases with addition of simulated images for training. In case of baseline, the performance increases by 2.6%, and by 4.4% for MSC-RCNN. This boost in performance can be attributed to increase in AP75 metric values compared to it's counter parts in real domain. Further, the detailed comparison of the model performance in terms of different scales of AP as indicated by APs, APm, APl metrics highlights the multi-scale segmentation performance of MSC-RCNN. The APs, APm, APl corresponds to the AP value for small, medium and large instances. The pixels corresponding to each of these instance is same as the default values used in [22].

Result on NWPU VHR-10 [6]: The proposed model performance is compared with other state-of-the-art models as shown in Table III. It can be observed that the proposed MSC-RCNN outperforms the other state-of-the-art models in terms of overall and AP75 metric values. This boost in performance is less compared to the one observed in the proposed slant angled imagery - SlantSim dataset. Hence, this highlights the MSC-RCNN superiority in detecting multi-scalar objects in slant angled imagery as well as the detection of small scaled and objects of different sizes as observed in overhead imagery.

Model	AP	AP50	AP75
ARENET[42]	64.8	93.2	71.5
Cascade Mask R-CNN[21]	59.9	90.0	64.9
Precise Mask R-CNN[34]	64.8	93.8	73.2
BMRSS[33]	65.2	94.9	72.1
SS-PANet[44]	65.9	94.2	76.7
MSC-RCNN	67.4	92	77.1

TABLE III: Instance segmentation performance on NWPU VHR 10.

VI. Ablation Study

A. Sim2Real adaptation

To experiment the extent of Sim2Real adaptation, the number of instances in simulated and real domain was varied. The number of real instances was decreased to 478 for fighter jet, 197 for helicopter, 1 for carrier plane and 47 for passenger plane. Whereas, the number of instances in

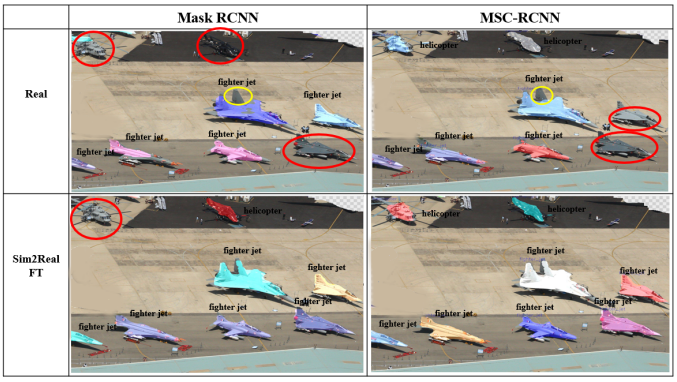

Fig. 5: Qualitative results of test data. The missed predictions are encircled.

Model	AP	AP50	AP75
MSC-RCNN (with GN)	50.3	87.6	64.2
MSC-RCNN (without GN)	52	92	62.8
Mask R-CNN	44.7	87.1	31.9

TABLE IV: Results of MSC-RCNN model with respect to Sim2Real domain adaptation.

Model	Dataset	AP	AP50	AP75
MSC-RCNN(with BN)	Real	33.4	80.2	19.8
MSC-RCNN(without GN)	Real	46.9	84.7	56.8
MSC-RCNN(with BN)	FT	49	88.6	58
MSC-RCNN(without GN)	FT	51.4	92.6	62.2

TABLE V: Results of MSC-RCNN model without GN for different experiments.

simulated domain was increased to 2249 in fighter jet, 702 in helicopter, 483 in carrier plane, 1276 in passenger plane class.

The results were evaluated for the proposed MSC-RCNN with and without GN fusion in it's backbone. The Table IV represents the evaluation results. It can be observed that, the performance as good as a model that was trained only on real images was achieved (refer to Table I). In addition, the following learning can be attributed to a single shot learning in case of carrier plane and few shot learning learning from real to simulated domain in case of other classes.

B. Group Normalisation

The addition of GN during the concatenation of FPN features makes sure that features are effectively concatenated as mentioned in Section 4.4. The same was proved by training the a) models without GN layer, b) inclusion of Batch Normalisation (BN) instead of GN after the concatenation. From Table V it can be observed that the inclusion of GN layers have greatly improved the performance and the localisation ability of the model by boosting the AP75 scores.

C. Parallel dilated FPN backbones

The number of parallel FPN backbones consisting of dilated convolutions was varied, and their performance was tabulated as shown in Table VI. MSC-RCNN3 and MSC-RCNN4 refer to 2 and 3 parallel dilated FPN backbones respectively. The arrangement of dilated convolutions in

Real					FT		
Params	Model	AP	AP50	AP75	AP	AP50	AP75
66.61	MSC-RCNN	51.4	89.9	65.4	55.8	91	64.7
70.00	MSC-RCNN3	50.2	88.9	65.3	52.1	89.6	65.5
74.12	MSC-RCNN4	48.2	87.3	59.7	51.8	86.8	61.9
66.71	Mask R-CNN+ACP[32]	49.7	89.4	63.2	51.3	90.1	64.3
73.67	Mask R-CNN+ASPP[5]	50.8	89	64.9	52.7	90.5	65.3

TABLE VI: Model performance with respect to parallel dilated convolutions.

these FPNs was the same as the one mentioned in section IV. With the inclusion of every additional dilated convolution FPN backbone, the model size increased by 5%. From Table VI, it can be observed that the best performance was achieved while using a network consisting of only 2 FPN backbones; hence, it was chosen as the final model for the proposed work. Furthermore, a comparative study with respect to existing multi-scale dilated convolution frameworks such as ACP[32] and ASPP[5] was performed. It can be observed that the proposed MSC-RCNN network outperforms these models in both real and Sim2Real FT domains with a relatively lesser model size.

VII. Conclusions

In this paper, a novel MSC-RCNN architecture is proposed, specifically designed to address the challenges of occlusions and varying object scales in slant-angle aerial imagery. Furthermore, SlantSim dataset, a novel synthetic slant-angle aerial imagery dataset, generated using Unreal Engine, is proposed to address the limitations of real-world data availability. The findings demonstrate that incorporating this synthetic data for Sim2Real transfer learning significantly reduces reliance on real data while maintaining high mean Average Precision (mAP) scores. Compared to the baseline Mask R-CNN, the proposed approach achieves a remarkable 7.6% improvement in instance segmentation accuracy with only a 6% increase in model size. This advancement paves the way for enhanced drone-based aerial surveillance in various applications. Future work includes expanding the dataset with additional object classes and plugging the proposed combined FPN backbone into other R-CNN networks such as Cascade Mask R-CNN and the Transformers, and evaluate their performances.

REFERENCES

[1] Jefferson Silva Almeida, Chenxi Huang, Fabrício Gonzalez Nogueira, Surbhi Bhatia, and Victor Hugo C. de Albuquerque. Edgefiresmoke: A novel lightweight cnn model for real-time video fire–smoke detection. *IEEE Transactions on Industrial Informatics*, 18(11):7889–7898, 2022.

[2] Adrian Boguszewski, Dominik Batorski, Natalia Ziemba-Jankowska, Tomasz Dziedzic, and Anna Zambrzycka. Landcover. ai: Dataset for automatic mapping of buildings, woodlands, water and roads from aerial imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1102–1110, 2021.

[3] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: High quality object detection and instance segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(5):1483–1498, 2021.

[4] Yaru Cao, Zhijian He, Lujia Wang, Wenguan Wang, Yixuan Yuan, Dingwen Zhang, Jinglin Zhang, Pengfei Zhu, Luc Van Gool, Junwei Han, et al. Visdrone-det2021: The vision meets drone object detection challenge results. In *Proceedings of the IEEE/CVF International conference on computer vision*, pages 2847–2854, 2021.

[5] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017.

[6] Gong Cheng and Junwei Han. A survey on object detection in optical remote sensing images. *ISPRS journal of photogrammetry and remote sensing*, 117:11–28, 2016.

[7] Ruggero Donida Labati, Angelo Genovese, Vincenzo Piuri, and Fabio Scotti. Wildfire smoke detection using computational intelligence techniques enhanced with synthetic smoke plume generation. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 43(4):1003–1012, 2013.

[8] Dawei Du, Yuankai Qi, Hongyang Yu, Yifan Yang, Kaiwen Duan, Guorong Li, Weigang Zhang, Qingming Huang, and Qi Tian. The unmanned aerial vehicle benchmark: Object detection and tracking. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[9] Epic Games. Unreal engine.

[10] Ross Girshick. Fast r-cnn. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1440–1448, 2015.

[11] Abdul Mueed Hafiz and Ghulam Mohiuddin Bhat. A survey on instance segmentation: state of the art. *International journal of multimedia information retrieval*, 9(3):171–189, 2020.

[12] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[14] Dániel Horváth, Gábor Erdős, Zoltán Istenes, Tomáš Horváth, and Sándor Földi. Object detection using sim2real domain randomization for robotic applications. *IEEE Transactions on Robotics*, 39(2):1225–1243, 2022.

[15] Meng-Ru Hsieh, Yen-Liang Lin, and Winston H Hsu. Drone-based object counting by spatially regularized regional proposal network. In *Proceedings of the IEEE international conference on computer vision*, pages 4145–4153, 2017.

[16] Zhaojin Huang, Lichao Huang, Yongchao Gong, Chang Huang, and Xinggang Wang. Mask scoring r-cnn. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[17] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. *Advances in neural information processing systems*, 28, 2015.

[18] Josy John, K Harikumar, J Senthilnath, and Suresh Sundaram. An efficient approach with dynamic multiswarm of uavs for forest fire-fighting. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2024.

[19] Konstantinos Kanistras, Goncalo Martins, Matthew J. Rutherford, and Kimon P. Valavanis. A survey of unmanned aerial vehicles (uavs) for traffic monitoring. In *2013 International Conference on Unmanned Aircraft Systems (ICUAS)*, pages 221–234, 2013.

[20] Alexander Kirillov, Yuxin Wu, Kaiming He, and Ross Girshick. Pointrend: Image segmentation as rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[21] Tsung-Yi Lin, Piotr Dollar, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[22] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[23] Xiaomin Lin, Cheng Liu, Allen Pattillo, Miao Yu, and Yiannis Aloimonous. Seadronesim: Simulation of aerial images for detection of objects above water. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 216–223, 2023.

[24] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 21–37. Springer, 2016.

[25] Zikun Liu, Liu Yuan, Lubin Weng, and Yiping Yang. A high resolution optical satellite image dataset for ship recognition and some new baselines. In *ICPRAM*, pages 324–331, 2017.

[26] Edward M Mikhail, James S Bethel, and J Chris McGlone. *Introduction to modern photogrammetry*. John Wiley & Sons, 2001.

[27] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[28] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

[29] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 102–118. Springer, 2016.

[30] Duccio Rocchini, Markus Metz, Alessandro Frigeri, Luca Delucchi, Matteo Marcantonio, and Markus Neteler. Robust rectification of aerial photographs in an open source environment. *Computers & geosciences*, 39:145–151, 2012.

[31] Jacob Shermeyer, Thomas Hossler, Adam Van Etten, Daniel Hogan, Ryan Lewis, and Daeil Kim. Rareplanes: Synthetic data takes flight. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 207–217, 2021.

[32] Zequn Sun, Chunning Meng, Jierong Cheng, Zhiqing Zhang, and Shengjiang Chang. A multi-scale feature pyramid network for detection and instance segmentation of marine ships in sar images. *Remote Sensing*, 14(24):6312, 2022.

[33] Sumanth Udupa, Prajwal Gurunath, Aniruddh Sikdar, and Suresh Sundaram. Mrfp: Learning generalizable semantic segmentation from sim-2-real with multi-resolution feature perturbation. *arXiv preprint arXiv:2311.18331*, 2023.

[34] Shridhar Velhal, Suresh Sundaram, and Narasimhan Sundararajan. A decentralized multirobot spatiotemporal multitask assignment approach for perimeter defense. *IEEE Transactions on Robotics*, 38(5):3085–3096, 2022.

[35] Syed Waqas Zamir, Aditya Arora, Akshita Gupta, Salman Khan, Guolei Sun, Fahad Shahbaz Khan, Fan Zhu, Ling Shao, Gui-Song Xia, and Xiang Bai. isaid: A large-scale dataset for instance segmentation in aerial images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 28–37, 2019.

[36] Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

[37] Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. Dota: A large-scale dataset for object detection in aerial images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3974–3983, 2018.

[38] Gui-Song Xia, Jingwen Hu, Fan Hu, Baoguang Shi, Xiang Bai, Yanfei Zhong, Liangpei Zhang, and Xiaoqiang Lu. Aid: A benchmark data set for performance evaluation of aerial scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(7):3965–3981, 2017.

[39] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015.

[40] Junpeng Zhang, Xiuping Jia, Jiankun Hu, and Kun Tan. Moving vehi-

cle detection for remote sensing video surveillance with nonstationary satellite platform. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):5185–5198, 2022.

[41] Zhitao Zhao, Ping Tang, Lijun Zhao, and Zheng Zhang. Few-shot object detection of remote sensing images via two-stage fine-tuning. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2022.

[42] Zhong-Qiu Zhao, Peng Zheng, Shou-tao Xu, and Xindong Wu. Object detection with deep learning: A review. *IEEE transactions on neural networks and learning systems*, 30(11):3212–3232, 2019.