AUTH-PERSONS: A DATASET FOR DETECTING HUMANS IN CROWDS FROM AERIAL VIEWS

Charalampos Symeonidis, Ioannis Mademlis, Ioannis Pitas and Nikos Nikolaidis

Department of Informatics, Aristotle University of Thessaloniki, Thessaloniki, Greece
Email: {charsyme, imademlis, pitas, nnik}@csd.auth.gr

ABSTRACT

Recent advances in artificial intelligence, control and sensing technologies have facilitated the development of autonomous Unmanned Aerial Vehicles (UAVs). Detecting humans from video input captured on-the-fly from UAVs is a critical task for ensuring flight safety, mostly handled with lightweight Deep Neural Networks (DNNs). However the detection of individual people in the case of dense crowds and/or distribution shifts (i.e., significant visual differences between the training and the test sets) is still very challenging. This paper presents AUTH-Persons, a new, annotated, publicly available video dataset, that consists of both real and synthetic footage, suitable for training and evaluating aerial-view person detection algorithms. The synthetic data were collected from 8 visually distinct photorealistic outdoor environments and they mostly contain scenes with crowded areas, where heavy occlusions and high person densities pose challenges to common detectors. This dataset is employed to evaluate the generalization performance of various state-of-the-art detection frameworks, by testing them on environments that are visually distinct from those they have been trained on. Finally, given that Non-Maximum Suppression (NMS) methods at the end of person detection pipelines typically suffer in crowded scenes, the performance of various NMS algorithms is also compared in AUTH-Persons.

Index Terms— person detection, Unmanned Aerial Vehicles, synthetic data generation, Non-Maximum Suppression

1. INTRODUCTION

Recent advances have led to an unprecedented popularization of Unmanned Aerial Vehicles (UAVs, or "drones") during the last decade. Drones have proven useful for many civilian and military applications, such as search and rescue operations, surveillance, inspection, mapping [1], wildlife monitoring, crowd monitoring/management, precision agriculture, or aerial media production [2] [3] [4] [5]. Gradual increases in UAV cognitive autonomy have made flight safety a critical issue, due to the hazard drones potentially pose in case of malfunction. Safety during UAV interactions with the environment must be particularly ensured when the vehicle is operating near humans. Autonomous UAVs should be able to visually detect people with a high-level of precision from various aerial views [6]. This task poses challenges due to the small size of objects/persons (especially in high flight altitudes), as well as due to unforeseen and wide-ranging variations in illumination, camera orientation, etc. Problems are exacerbated when UAVs must detect the presence of individuals within crowded areas.

A typical object detector, employed in real-life conditions, must be trained on multiple datasets in order to improve its generalization abilities and ensure its robustness during actual deployment. In recent years, several real-world and synthetic datasets have been proposed to tackle the problem of detecting humans from aerial images/video frames. *Stanford Drone Dataset* [7] is a large-scale video dataset consisting of 60 annotated videos for detection and tracking of various object classes, including pedestrians. The *Okutama-Action* [8] dataset, mainly developed for concurrent human action detection, can also be employed for person detection. It consists of 43 minute-long fully-annotated sequences with the corresponding bounding-boxes/Regions-of-Interest (ROIs) and 12 action classes. The *VisDrone dataset* [9] consists of 263 video clips with 179,264 frames and additional 10,209 static images. The videos/images were acquired by various drone platforms, including the DJI Mavic Phantom series, and depict various scenarios across 14 cities in China. The dataset is suitable for image and video object detection, as well as for single/multi-object detection tracking.

In the general person detection task, the use of synthetic datasets has recently gathered pace since they can viably replace or augment real-world training data. In [10], 3D human models rendered on random backgrounds are employed to train a pedestrian detector. In a similar fashion, [11] inserts realistic DNN-generated 3D human models into existing natural background images, while trying to select appropriate scale and insertion locations. In [12], a graphical simulator is proposed which can automatically generate datasets for pedestrian and crowd analysis. Experimental evaluation

The AUTH-Persons dataset is available at: `https://aiia.csd.auth.gr/open-multidrone-datasets/`. This research has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreements No 731667 (MULTIDRONE) and No 871449 (OpenDR).

on crowd estimation showed that DNN models which were pretrained on a synthetic dataset and later finetuned with the real-world dataset, outperformed models trained exclusively with real-world data.

A typical case where person detection methods may drastically fail to perform is when they operate on images depicting dense crowds [13]. Non-Maximum Suppression (NMS), which is a common post-processing step typically placed at the end of the overall person detection pipeline, suffers especially in crowded scenes. NMS methods prune the number of overlapping detected raw candidate ROIs generated by a detector, in order to assign a single and spatially accurate detection to each object. The de facto standard in NMS for object detection is Greedy-NMS [14]. It selects high-scoring detections and deletes less confident neighbours, since they most likely cover the same object. An Intersection-over-Union (IOU) threshold determines which less confident neighboring detections are suppressed. Modern alternatives include Soft-NMS [15], where a rescoring function decreases the score of neighboring less confident detections, instead of completely eliminating them, achieving better precision and recall rates. GossipNet [16] is a DNN designed to perform NMS, by processing the coordinates and scores of the detections. Overall, it jointly analyzes all detections in the image, so as not to directly prune them, but to rescore them. GossipNet was modified in [17] for the specific case of person detection from aerial views, so as to jointly process visual appearance and geometric properties of candidate ROIs. More recently, [18] proposed *Distance-IoU* (DIoU), a new metric which can replace the typical IoU metric in Greedy-NMS, by also considering the distance between the centers of two neighboring detections. Alternatively, Cluster-NMS was proposed in [19], i.e., a method where NMS is performed by implicitly clustering candidate detections, achieving very fast inference runtimes. Finally, a DNN-based NMS method called *Seq2Seq-NMS* [20] achieved top person detection results by assuming a sequence-to-sequence formulation of the NMS problem, exploiting the Multihead Scale-Dot Product Attention mechanism and jointly processing both geometric and visual properties of the input candidate detections.

This paper introduces *AUTH-Persons*, a large-scale dataset containing videos that depict human crowds from an aerial point of view and is suitable for training/evaluating relevant person detection methods. The paper makes the following contributions:

- AUTH-Persons is presented and made publicly available. It contains both synthetic videos, from diverse and realistic landscape environments, and real-world videos collected at the campus of the Aristotle University of Thessaloniki, Greece. In both cases, the footage was collected using UAVs performing flights at various altitudes, while crowds of people were present on the ground. All video frames are annotated with 2D bounding boxes.
- AUTH-Persons was constructed with the explicit aim to study person detection performance in the presence of distribution shifts between the training and the test set in crowded scenes. This has barely been explored in the literature. Thus, experimental evaluation of multiple DNN-based person detectors is conducted on AUTH-Persons in a manner that allows us to assess their ability to generalize to environments that are visually distinct from those they have been trained with.
- To complement this study, a lightweight version of YOLOv4 [21] is employed to evaluate the performance of several NMS methods at the end of the person detection pipeline, in an attempt to examine the negative impact of this data distribution shift on them. Such a study is of particular interest, because NMS algorithms are particularly susceptible to performance degradation in crowded scenes.

The dataset is available at: `https://aiia.csd.auth.gr/open-multidrone-datasets/`.

2. DATASET DESCRIPTION

AUTH-Persons is a UAV video dataset containing 53 videos, summing to footage with a total duration of 37.31 minutes. It is suitable for training and evaluating methods related to the person detection task. Overall, 4 of those videos were collected from a *DJI Phantom 4* while performing flights in the campus of the Aristotle University of Thessaloniki, Greece. The resolution of real-world video frames is 3840×2160 pixels. The remaining videos were collected in virtual environments, using AirSim [22] and a set of environments we designed on Unreal Engine 4[1]. We designed 8 environments, aiming to realistically simulate various environmental conditions (e.g., snow, fog, etc.) in rural landscapes. All environments were populated with a large number of humans and obstacles (e.g., trees, structures, etc.), in order to achieve a high level of occlusions. The footage was collected from a virtual UAV, while orbiting around in various altitudes. The resolution of these synthetic video frames is 1280×720 pixels. Video frames from *AUTH-Persons* are depicted in Figure 1. The frame-rate of all videos is set to 30 fps. The average number of people depicted in each frame is 14.79. 2D bounding boxes of humans are provided as annotations for each frame. Details about the structure of the dataset are provided in Table 1.

3. EXPERIMENTAL EVALUATION

The evaluation conducted on AUTH-Persons, using recent DNN-based object detectors and NMS methods, is presented here. Exploiting the diversity of the dataset, AUTH-Persons was split so that the test set contains environments that are visually distinct from those included in the training set. Thus, environments I-VII were selected for training and the rest for testing. This setup simulates the case where a person detector,

[1] `https://www.unrealengine.com/en-US/`

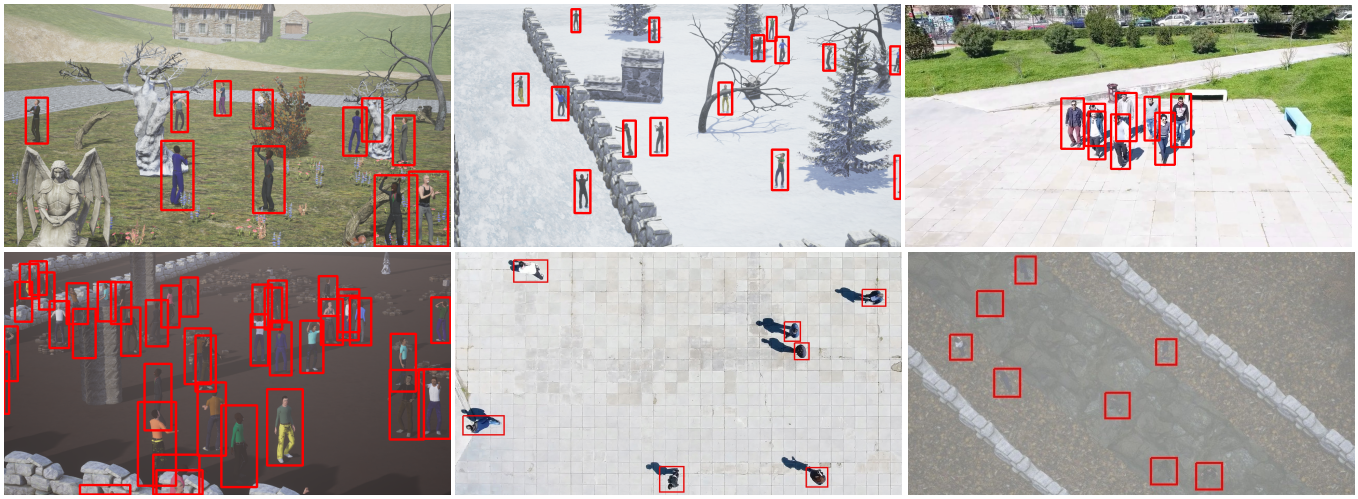

Fig. 1: Video frames from the *AUTH-Persons* dataset, depicting both real and synthetic environments. The ground-truth bounding boxes surrounding visible humans are depicted in red.

Table 1: Structure of the AUTH-Persons Dataset.

Env. ID	Synthetic/ Real-World	Num. of Videos	Duration [mm:ss]	Resolution	Aver. Num. of Persons per Frame
I	S	7	03:33	1280×720	13.50
II	S	6	03:34	1280×720	13.95
III	S	6	04:29	1280×720 +	16.14
IV	S	4	03:14	1280×720	18.15
V	S	6	03:27	1280×720	31.34
VI	S	10	05:19	1280×720	10.75
VII	R	1	02:22	3840×2160	8.69
VIII	S	3	02:12	1280×720	16.93
IX	S	7	05:41	1280×720	15.67
X	R	3	03:40	3840×2160	3.9
–	–	53	37:31	–	14.79

embedded in an autonomous UAV, is deployed on an unseen environment. The goal was to measure the impact of such a data distribution shift on the detector's precision, as well as on the performance of its integrated DNN-based NMS method.

Evaluation of Person Detection Methods: First, we report results of three person detectors: the Single Shot Detector (SSD) [23], YOLOv3 [24] and YOLOv4-tiny [21]. Various Convolutional Neural Networks (CNNs) are employed as backbone feature extractors. Both the training and the test set were constructed by sampling 1 out of 10 consecutive video frames. All detectors were trained for 15 epochs. Their learning rate was initially set to 5×10^{-4} and it was decreased twice by multiplying it with 0.1 at epochs 10 and 13, respectively. The remaining training hyperparameters were adjusted in the best possible manner, so as to achieve a fair comparison. Greedy-NMS with a 0.6 IoU threshold was applied as the last step on all detectors. The results on both sets are reported in Table 2.

All detectors exhibit a significant precision drop of at least 8% in $AP_{0.5}$ in the test set, compared to the training set. The obvious explanation is the distribution shift between the training and the test data, due to the visually different environments depicted in those two sets. The best precision rates in the test set were achieved by YOLOv4-tiny, although it did not achieve top precision in the training set. Overall, YOLOv4-tiny seems to be the most robust method.

Table 2: Person Detection Evaluation.

Detector	Training Set		Test Set	
	$AP_{0.5}$	$AP^{0.95}_{0.5}$	$AP_{0.5}$	$AP^{0.95}_{0.5}$
SSD-512 (VGG16_atrous)	88.6%	52.3%	76.7%	40.4%
SSD-512 (ResNet50)	85.8%	47.8%	73.8%	36.7%
SSD-512 (MobileNetV1-1.0)	84.3%	43.7%	68.4%	31.1%
YOLOv3-512 (DarkNet53)	**94.6%**	61.5%	81.9%	**48.4%**
YOLOv3-512 (MobileNetV1-1.0)	92.3%	55.9%	77.4%	41.5%
YOLOv3-800 (MobileNetV1-1.0)	**94.6%**	**61.9%**	79.8%	47.2%
YOLOv4-tiny-608 (CSPDarknet53-tiny)	93.7%	56.3%	**85.0%**	47.3%

Evaluation of NMS Methods: YOLOv4-tiny was selected as the main person detector for the evaluation of the NMS methods. In this setup we compare the performance of the recently proposed Seq2Seq-NMS [20] and a wealth of other state-of-the-art NMS methods. The second competing method is a baseline Greedy-NMS approach running on CPU. The third is TorchVision's [2] Greedy-NMS implemented to run very fast on GPUs. Additionally, the non-neural approach Soft-NMS [15] was tested, using both the proposed linear and the Gaussian weighting functions (referred to as Soft-NMS_L and Soft-NMS_G, respectively). The method was executed on CPU. Furthermore, several variants of the more recent Cluster-NMS [19] non-neural approach were selected for comparison purposes. In what follows, the term Cluster-NMS_S is used to imply the case where the score penalty mechanism is used, and Cluster-NMS_D the scenario where the normalized central point distance is added. In the latter case, the method is equivalent to DIoU-NMS [18]. Moreover, the term Cluster-NMS_{S+D} is used when both of these mechanisms are utilized. Finally, Cluster-NMS_{S+D+W} indicates the case where a weighted strategy is applied. The last selected approach is GossipNet [16], a neural NMS method achieving good precision.

[2] https://pytorch.org/vision/stable/ops.html#torchvision.ops.nms

The vast majority of NMS methods operate by only analyzing geometric relations between raw candidate detection ROIs in an image. Very few, such as Seq2Seq-NMS, exploit also visual appearance information. However, the visual data distribution shift between the training and the test samples may disproportionately affect in a negative manner those DNN-based NMS methods which do exploit appearance-based features in comparison to those who do not. Thus, Seq2Seq-NMS was evaluated in two variants: a) the vanilla Seq2Seq-NMS [20], and b) a trivial variant Seq2Seq-NMS$_{geom}$ which only exploits geometry-based features without considering visual appearance. Seq2Seq-NMS$_{geom}$ was implemented by simply feeding the DNN a zero vector for each ROI, as a dummy appearance-based feature.

Table 3: Comparison of different NMS methods on AUTH-Persons dataset using detections from YOLOv4-tiny [21].

Method	**Device**	**Test set**		**Average Inference Time (ms)**
		AP$_{0.5}$	**AP**$_{0.5}^{0.95}$	
Greedy-NMS IoU>0.5	CPU	84.7%	46.8%	1.8
Greedy-NMS IoU>0.6	CPU	85.0%	47.3%	1.9
Greedy-NMS IoU>0.7	GPU	83.9%	47.5%	2.1
Original NMS IoU>0.5 TorchVision	GPU	84.7%	46.9%	0.4
Original NMS IoU>0.6 TorchVision	GPU	84.9%	47.3%	0.4
Original NMS IoU>0.7 TorchVision	GPU	83.8%	47.5%	0.4
Soft-NMS$_{\text{L}}$	CPU	84.5%	**48.0%**	2.5
Soft-NMS$_{\text{G}}$	CPU	84.6%	47.9%	1.9
Cluster-NMS	GPU	84.9%	47.4%	1.3
Cluster-NMS$_{\text{S}}$	GPU	84.3%	47.8%	1.6
Cluster-NMS$_{\text{D}}$	GPU	85.0%	47.2%	1.9
Cluster-NMS$_{\text{S+D}}$	GPU	84.5%	47.9%	2.0
Cluster-NMS$_{\text{S+D+W}}$	GPU	84.4%	47.4%	15.6
GossipNet	GPU	85.4%	47.4%	5.6
Seq2Seq-NMS	GPU	85.2%	46.9%	15.8
Seq2Seq-NMS$_{geom}$	GPU	**85.5%**	**48.0%**	15.8

The hyperparameters of all competing NMS methods were set according to the original respective papers. In order to account for the extremely large number of raw detections in several frames, we first apply TorchVision NMS with the relaxed 0.8 IoU threshold on all deployed methods as a typical preprocessing step. A similar scheme was also employed in the training and the testing phases of GossipNet and Seq2Seq-NMS. All experiments were performed on a PC using an Intel Core i7-9700 CPU and an NVIDIA GeForce RTX 2070 GPU with 8GB of memory, both for training and inference.

NMS evaluation results on AUTH-Persons are provided in Table 3. Vanilla Seq2Seq-NMS, which exploits both appearance-based and geometry-based features, achieves an $AP_{0.5}$ of 85.2%, which is the second highest after GossipNet. Seq2Seq-NMS$_{geom}$ improves upon vanilla Seq2Seq-NMS in $AP_{0.5}$ and $AP_{0.5}^{0.95}$ by +0.3% and +1.1%, respectively, thus rendering it the overall best NMS method for AUTH-Persons. Notably, Soft-NMS$_L$ also achieves a top $AP_{0.5}^{0.95}$ score, equal to that of Seq2Seq-NMS$_{geom}$. The non-negligible improvement of Seq2Seq-NMS$_{geom}$ over vanilla Seq2Seq-NMS confirms our intuition that NMS methods exploiting appearance-based features suffer more in the presence of distribution shifts, compared to methods that only exploit geometry-based features.

4. CONCLUSIONS

Detecting humans on video footage captured on-the-fly by UAVs is a challenging, yet critical task for ensuring flight safety in the case of autonomous drones. However, aerial detection of individual people in crowds under the presence of distribution shifts is still very challenging. This paper presented AUTH-Persons in order to facilitate relevant research. It is a new, publicly available video dataset that consists of both real and synthetic footage, summing approximately to a duration of 37 minutes. The dataset is suitable for training and evaluating aerial-view person detection algorithms. The synthetic data were collected from 8 visually distinct photorealistic outdoor environments and they mostly contain scenes with crowded areas, where heavy occlusions and high person densities pose challenges to common detectors. The generalization performance of various state-of-the-art DNN-based object detectors was evaluated on AUTH-Persons, by testing them on environments that are visually distinct from those they have been trained on. YOLOv4-tiny was empirically shown to be the most robust person detector. Finally, given that Non-Maximum Suppression (NMS) methods at the end of the person detection pipeline suffer in crowded scenes, they were compared with respect to the degree that train-to-test distribution shifts affect them in such settings. Among DNN-based free-rescoring NMS algorithms, Seq2Seq-NMS and Soft-NMS achieved top precision, while appearance-based features are more likely to lead to NMS performance degradation under distribution shift, compared to purely geometry-based ones.

5. REFERENCES

[1] F. Nex and F. Remondino, "UAV for 3D mapping applications: a review," *Applied Geomatics*, vol. 6, no. 1, pp. 1–15, 2014.

[2] I. Mademlis, V. Mygdalis, N. Nikolaidis, M. Montagnuolo, F. Negro, A. Messina, and I. Pitas, "High-level multiple-UAV cinematography tools for covering outdoor events," *IEEE Transactions on Broadcasting*, vol. 65, no. 3, pp. 627–635, 2019.

[3] I. Karakostas, I. Mademlis, N. Nikolaidis, and I. Pitas, "Shot type constraints in UAV cinematography for autonomous target tracking," *Elsevier Information Sciences*, vol. 506, pp. 273–294, 2020.

[4] I. Karakostas, I. Mademlis, N. Nikolaidis, and I. Pitas, "Shot type feasibility in autonomous UAV cinematography," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.

[5] I. Mademlis et al., "A multiple-UAV architecture for autonomous media production," *Springer Multimedia Tools and Applications*, pp. 1–30, 2022.

[6] C. Symeonidis, E. Kakaletsis, I. Mademlis, N. Nikolaidis, A. Tefas, and I. Pitas, "Vision-based UAV safe landing exploiting lightweight Deep Neural Networks," in *Proceedings of the International Conference on Image and Graphics Processing (ICIGP)*, 2021.

[7] A. Robicquet, A. Sadeghian, A. Alahi, and S. Savarese, "Learning social etiquette: Human trajectory understanding in crowded scenes," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016.

[8] M. Barekatain, M. Martí, H.F. Shih, S. Murray, K. Nakayama, Y. Matsuo, and H. Prendinger, "Okutama-action: An aerial view video dataset for concurrent human action detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2017, pp. 2153–2160.

[9] 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.

[10] L. Pishchulin, A. Jain, C. Wojek, M. Andriluka, T. Thormählen, and B. Schiele, "Learning people detection models from few training samples," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[11] C. Symeonidis, P. Nousi, P. Tosidis, K. Tsampazis, N. Passalis, A. Tefas, and N. Nikolaidis, "Efficient realistic data generation framework leveraging deep learning-based human digitization," in *Proceedings of the 22nd Engineering Applications of Neural Networks Conference*, 2021.

[12] A. Khadka, P. Remagnino, and V. Argyriou, "Synthetic crowd and pedestrian generator for deep learning problems," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

[13] L. Songtao, H. Di, and W. Yunhong, "Adaptive nms: Refining pedestrian detection in a crowd," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[14] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1627–1645, 2010.

[15] N. Bodla, B. Singh, R. Chellappa, and L. S. Davis, "Soft-NMS: Improving object detection with one line of code," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[16] J. Hosang, R. Benenson, and B. Schiele, "Learning Non-Maximum Suppression," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[17] C. Symeonidis, I. Mademlis, N. Nikolaidis, and I. Pitas, "Improving neural Non-Maximum Suppression for object detection by exploiting interest-point detectors," in *Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, 2019.

[18] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, and D. Ren, "Distance-IoU loss: Faster and better learning for bounding box regression," *Association for the Advancement of Artificial Intelligence (AAAI)*, vol. 34, no. 7, pp. 12993–13000, 2020.

[19] Z. Zheng, P. Wang, D. Ren, W. Liu, R. Ye, Q. Hu, and W. Zuo, "Enhancing geometric factors in model learning and inference for object detection and instance segmentation," *ArXiv*, vol. abs/2005.03572, 2020.

[20] C. Symeonidis, I. Mademlis, I. Pitas, and N. Nikolaidis, "Neural attention-driven non-maximum suppression for person detection," *TechRxiv*, vol. 10.36227/techrxiv.16940275.v1, 2021.

[21] J. Zicong, Z. Liquan, L. Shuaiyang, and J. Yanfei, "Real-time object detection method based on improved yolov4-tiny," *ArXiv*, vol. abs/2011.04244, 2020.

[22] S. Shah, Debadeepta Dey, Chris Lovett, and Ashish Kapoor, "Airsim: High-fidelity visual and physical simulation for autonomous vehicles," in *Field and Service Robotics (FSR)*, 2017.

[23] W. Liu, D. Anguelov, D.and Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2016.

[24] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, real-time object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.