A UAV Video Data Generation Framework for Improved Robustness of UAV Detection Methods

Charalampos Symeonidis
Department of Informatics
Aristotle University of Thessaloniki
Thessaloniki, Greece
charsyme@csd.auth.gr

Charalampos Anastasiadis
Department of Informatics
Aristotle University of Thessaloniki
Thessaloniki, Greece
xaranastasiadis@gmail.com

Nikos Nikolaidis
Department of Informatics
Aristotle University of Thessaloniki
Thessaloniki, Greece
nnik@csd.auth.gr

***Abstract*—Recent advances have facilitated the development and popularization of Unmanned Aerial Vehicles (UAVs) that can operate semi or fully autonomously. The real-time, accurate visual detection of UAVs is crucial for various tasks and applications including surveillance (e.g., detecting UAVs flying over restricted areas such as airports) or multi-robot systems (e.g., a swarm of UAVs that need to cooperate and avoid collisions between swarm members in GPS-denied environments). The small target-to-image ratio and large similarity with other flying objects makes the visual detection of UAVs a challenging task. In addition, data distribution shifts can have a major negative impact to UAV detection frameworks, often trained on a wide variety of datasets to achieve an adequate level of robustness. As an attempt to mitigate the effect of these issues, we present a method that can generate realistic annotated video data depicting flying UAVs, using as input real background videos and 3D UAV models. The conducted experimental evaluation showed that the synthetic data are both challenging and realistic and that detectors trained on a combination of real-world and synthetic data, exhibit an improved generalization performance, achieving better precision rates when evaluated on real datasets that are visually distinct from the corresponding real training data.**

***Index Terms*—Drone/UAV Detection, Synthetic Data Generation**

I. Introduction

Technological progress has led to an increasing use of semi or fully autonomous Unmanned Aerial Vehicles (UAVs), with complex signal processing, computer vision and machine learning algorithms facilitating their operation. UAVs have proven useful for many civilian and military applications, such as precision agriculture, inspection, search and rescue operations, mapping [11], wildlife monitoring, crowd monitoring/management [19], or aerial media production [10]. Usually UAVs are equipped with a single or multiple cameras in order to capture visual information about their surroundings and take decisions.

Occasionally, the UAVs must be able to perceive the existence of other UAVs in the nearby airspace for tasks related to surveillance, flight safety (e.g., collision avoidance with other UAVs [9]), multi-robot cooperative missions (e.g., self-organising swarms of UAVs in GPS-denied environments [20]) or UAV cinematography (e.g. avoiding having a UAV entering the field of view of another UAV in events covered by multiple UAVs). To this end, a general visual object detector is often incorporated in their overall system, assigned with the task to detect UAVs along with other objects, for which information about their existence is essential in their operation. Visual UAV detection is also needed in ground surveillance systems, for example systems assigned the task of surveying a restricted airspace for unauthorized UAV presence. However detecting UAVs is a challenging task, requiring in certain cases a detector that operates explicitly for this task. Such major challenges are the difficulty of detecting UAVs in long distances due to their small size, color and shape variations of UAVs, poor weather (e.g., fog, rain) and illumination conditions, complex background when the UAV is filmed over an urban backdrop, and the visual similarities between UAVs and other aerial objects, such as birds or airplanes.

UAV detectors, deployed on real-world scenarios, need to be pretrained on a wide range of positive samples (e.g., various UAV models in multiple colors and views) and negative samples (e.g., birds, airplanes, background environments, etc.) to achieve an adequate level of robustness. In the recent literature, a significant number of datasets related to the UAV/drone detection task have been proposed. The *UAV Dataset: Amateur Unmanned Air Vehicle Detection* [3] comprises of more than 4000 annotated images. In most of them, the DJI Phantom is depicted. The image resolution varies between 300x168 pixels and 4K. The dataset contains also images with non-UAV objects. The *Real World Object Detection Dataset for Quadcopter Unmanned Aerial Vehicle Detection* [13] is a dataset consisting of 51446 annotated images for training and 5375 annotated images for testing. Its images were either collected from the Internet or recorded by the authors and have all been scaled to a resolution of 640×480. Along with the dataset, the authors present the performance of various detection algorithms on their dataset, as baselines for the development of more complicated detection systems. In addition, the authors in [16] presented a novel video UAV detection dataset containing 650 annotated infrared and RGB videos of UAVs, birds and airplanes and helicopters. The dataset also contains audio clips of the classes UAVs, helicopters and background noise. Finally, a large-scale, diverse dataset is the *UAV vs Bird Detection Challenge* [6] dataset. The dataset has been used at the corresponding challenge for evaluating the performance of algorithms in the task of detecting small

Fig. 1: Assets utilized from the proposed data generation method. *First Row:* Frames from videos captured from UAVs. *Second Row:* 3D models of various UAV models and birds, employed as distractors. (from left to right: DJI Phantom 3, Parrot AR 2.0 and a bird, each in two color variations.)

UAVs in image data. The challenge focuses on the ability of the detection methods to visually distinguish between UAVs and birds, particularly at large distances from the capturing camera. The training set consists of 77 videos, each comprising of 1384 frames on average. The test set consists of 14 videos for which no annotations are provided.

Apart from real-world datasets, a number of synthetic datasets, suitable for training object detection methods, or approaches for generating the corresponding images, have been proposed. In [14], the authors estimate the rendering parameters required to synthesize similar to real-world images, given a coarse 3D model of the target object. The authors in [5], have generated a synthetic dataset of 6k depth maps of UAVs, using AirSim simulator [15] in order to train a deep learning-based UAV detection model with the aim to be deployed on a dynamic obstacle avoidance method.

Although, as mentioned above, an increasing number of real-world datasets suitable for UAV detection have become available in recent years, most of them are relatively poor in terms of visual diversity, since they involve. for example, a single/small number of view-points [16], or depict a limited number of UAV models [3]. In this paper we propose an effective and low-cost visual data generation method for UAV detection that:

- is capable of reusing existing diverse aerial videos collected from camera-equipped flying UAVs and augmenting them with realistic 3D UAV models, thus eliminating the need to film videos,
- provides automatically-generated, detailed annotations,
- increases the robustness of UAV detectors, by providing realistic and diverse training data suitable to be employed for their training.

To complement this study, the state-of-the-art YOLOv4 [4] and YOLOv4-tiny [21] detectors are employed, in an attempt to examine whether our synthetic data a) are challenging and realistic and b) can have a positive impact on the performance of a detector when they are utilized for its training along with real-word data. The corresponding detectors were selected since they can effectively detect objects in various scales, including objects of very small scales in images of sufficient input resolutions and they can achieve fast inference runtimes. In addition, since YOLOv4-tiny is a lightweight detection method, it is extremely suitable for operating on embedded devices with limited computing and memory resources that are installed on autonomous systems such as UAVs [12] [17] [7].

II. Proposed Data Generation Method

A setting similar to [18] was adopted for the data generation process. The proposed method requires as input a) a set of aerial videos captured by flying UAVs, depicting various real-world environments, b) a set of 3D UAV models and c) a set of 3D models of various flying objects that are visually similar to UAVs, such as airplanes or birds. Aerial videos captured by having a UAV flying over different areas, videos collected from the Internet, or suitable videos collected from pre-existing datasets such as the *MultiDrone Dataset* [1] or *YouTube-8M* [2] can be employed for the data generation process. In addition, a wide-range of 3D models of UAVs and other flying objects can be easily acquired from the Web. Texture re-coloring of those 3D models can be employed to increase their visual variety. Images of such videos and 3D models are depicted in Fig 1. Using those assets, a virtual environment can be constructed, aiming to realistically simulate a scenario where a UAV is flying while being recorded by another UAV's camera.

The described environment was set up on Unreal Engine 4 [8], while AirSim [15] was used for issuing control commands to the UAV model. However other game or simulation engines can be employed for this data generation technique. More specifically, the environment consists of a "projection" rectangle, where the input videos are projected, and a virtual camera, properly placed at [0,0,0], with the z-axis being parallel to camera's optical axis in a right-hand coordinate system, while the camera's image plane is parallel to the projection rectangle. The virtual camera is calibrated so that the videos depicted in the rectangle are fully visible, i.e. they cover the camera field of view. In our setup we set the AirSim-

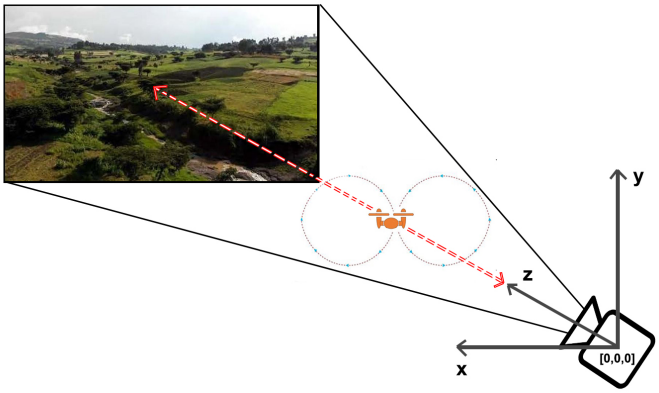

Fig. 2: Rough visualization of the data generation virtual environment.

controlled UAV model to move in a trajectory, where at any given time, it remains visible to the camera. This setup and the UAV model trajectory are depicted in Fig 2. In a simplified camera perspective model, the position of the UAV at any given time t is expressed in world-coordinates as:

$$x_t = \frac{wz_t}{4f}(\sin(\frac{t\pi}{d}) + 1),$$
$$y_t = \frac{hz_t}{4f}(\sin(\frac{t\pi}{d}) + 1),$$
$$z_t = z_{min} + (z_{max} - z_{min}) \begin{cases} \frac{t \pmod d}{d}, & \text{if } t \pmod d \leq 0.5 \\ 1 - \frac{t \pmod d}{d}, & \text{if } t \pmod d > 0.5 \end{cases}$$

where d is the time the UAV needs to complete a single traversal of its trajectory, z_{min} and z_{max} are the minimum and maximum coordinate values of the UAV's trajectory in z-axis, w and h are the width and height of the camera's image plane and f is the camera's focal length. In this formulation, the UAV follows an "eight-shaped" trajectory, in the xy plane while it simultaneously moves away and then back towards camera in the z axis. The other 3D assets, including birds or non-UAV aircrafts can perform similar or linear trajectories in the scene.

The annotations of the generated data, consist of the 2D bounding boxes of the depicted UAV model in each frame. The UAV must be encapsulated in a 3D bounding box during the data generation process. At each frame, recorded by the camera, the UAV's 2D bounding box is automatically generated based on the outer x, y image plane coordinates of the projection of the vertices of its 3D bounding box.

III. Experimental Evaluation

To evaluate the realism and the impact of the synthetic data on the performance of state-of-the-art deep learning object detectors, a dataset was generated based on the proposed method. Three UAV 3D models, namely the Parrot AR UAV 2.0, the DJI Mavic 2 pro and the DJI Phantom 3 were selected for the data generation process. As non-UAV aerial objects we used sparse flocks of animated birds colored in black, grey and white colors. Finally as backgrounds we selected 21 videos collected from cameras mounted on UAVs, depicting rural or urban environments. Overall, for each real-world background video, we generated a series of synthetically augmented video data, using variations regarding the depicted UAV model, its color, its projected size, and on whether the background behind the UAV depicted sky (almost uniform background) rural landscape (low complexity background) or buildings (highly complex background). The selected colors of each employed UAV model are depicted in Tab I. The dataset consists of 492 videos and was split into a training set, consisting of 300 videos, and a testing set with the remaining 192 videos. Frames of the generated dataset are illustrated in Fig. 3. Currently, we cannot make the dataset publicly available due to restrictions regarding the background videos. Once permissions are granted, we will make it available for scientific use.

TABLE I: Color variations of each UAV model employed in the data generation process.

UAV Models	UAV Colors
Parrot AR UAV 2.0	Black Grey White
DJI Mavic 2 Pro	Grey Black White Red
DJI Phantom 3	White Grey Black Yellow

In the first part of the experimental evaluation we examine whether our data are challenging for state-of-the-art detection frameworks. For this task, we trained a vanilla YOLOv4-tiny [21] model for 64 epochs. Then, the model was evaluated on the full test set, as well as on several of its subsets aiming to examine the impact of the depicted background environment and the depicted UAV's size, on the detector's performance. The results are shown on Tab. II.

TABLE II: Performance of YOLOv4-tiny on our synthetic dataset.

Test set	Average Precision		Average Recall
	$\mathbf{AP}_{0.5}$	$\mathbf{AP}_{0.5}^{0.95}$	$\mathbf{AR}_{0.5}^{0.95}$
UAV size: Large	95.0%	79.2%	82.9%
UAV size: Medium	96.4%	74.6%	78.7%
UAV size: Small	83.2%	56.7%	61.4%
Background: Buildings	89.0%	66.1%	70.3%
Background: Sky	92.3%	72.8%	75.9%
Background: Landscape	92.9%	67.6%	71.4%
Full Test Set	90.1%	67.1%	70.9%

Overall, the detector achieved good precision and recall rates in the full test set. Among the conducted evaluation variations, the detector was, as expected, mostly challenged when the UAV's depicted size was small, achieving an $AP_{0.5}$

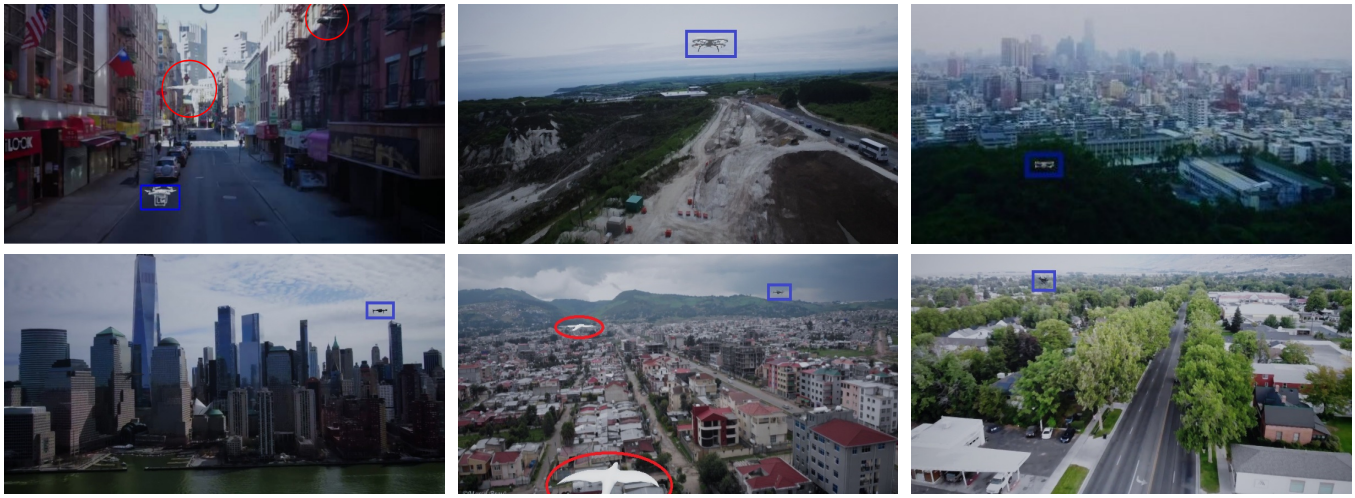

Fig. 3: Images from the synthetic video dataset for UAV detection. The UAVs are highlighted in blue boxes, while the birds, used for making the dataset more challenging, are highlighted in red circles.

of 83.2% and an $AR_{0.5}$ of 61.4% as well as when the real-world videos on which the UAV model was projected depicted urban environments (e.g., building, streets, etc.) achieving an $AP_{0.5}$ of 89.0% and an $AR_{0.5}^{0.95}$ of 70.3%.

TABLE III: Formulation of the datasets used in the experimental evaluation study. "R + S" refers to using real and synthetic data for training, while "R" refers to using only real data for training or testing.

Dataset acromym	Source	Training sets			Test set
		Number of images in R	Number of images in R+S	Synthetic to Real-World samples ratio	Number of images in R
Dataset 1	Data from [3]	3611	4426	22.6%	400
Dataset 2	Data from [6]	2040	2877	41.0%	252
Dataset 3	Data from [16]	4136	4399	6.4%	1432

To conduct the second part of our experimental evaluation, we employed three real-world UAV datasets presented in [3], [6] and [16]. A random training-test data split was performed in [6] and [16], since the first doesn't provide an annotated test set, and the last doesn't provide a formal data split. In addition, for those two video datasets, we sampled only a few frames from each video. For each dataset, we created two training sets. The first contains only its corresponding real-world data, while the second is augmented with a small amount of synthetic samples. The synthetic to real-world samples ratio varies from 6% to 41% between the three datasets. The synthetic samples in all three datasets were selected from the same subset of our dataset, using different sampling ratios. The test sets contain only real-world images. Information regarding the datasets, and their corresponding splits is provided in Tab. III.

Aiming to asses the level of realism of the generated synthetic data, we trained a vanilla YOLOv4 [4] model on the three separate datasets, whose training set contains mostly real-world data along with a small percentage of synthetic data. All detectors were trained for 64 epochs and tested on the same subset of our synthetic dataset. As shown on Tab. IV, the model trained on Dataset 2_{R+S} achieved a high $AP_{0.5}$ rate of 89.7%. This high precision rate, can be explained by the fact that the real-world training images are far more diverse in Dataset 2 compared to Datasets 1 and 3. More specifically, the images from [6], which comprise the real-world data of Dataset 2, depict a larger amount of diverse environments while also being collected from a greater variety of viewing angles, compared to [3] [16]. However, all models achieve good precision rates, demonstrating that the synthetic data are indeed visually similar to real-world data.

TABLE IV: Evaluation of YOLOv4, using real and synthetic data (R+S) for training and tested on synthetic data.

Training Set	$\mathbf{AP}_{0.5}$
Dataset 1 (R+S)	66.6%
Dataset 2 (R+S)	89.7%
Dataset 3 (R+S)	60.9%

Finally, we assess the impact of synthetic data used as training samples along with real-world data, on the performance of DL-methods tested on real-world data only. More specifically, two YOLOv4 models were trained for each of the three real-world UAV datasets. The first model was trained using only the real-world training set (R) of a UAV dataset, while the second was trained on the same set augmented with synthetic data (R+S). Both models were identical, in terms of parameters, and were trained for the same number of epochs. Then, both models were evaluated on the test set of the same dataset (intra-dataset), as well as on the test sets of the other two datasets (inter-dataset). Through the intra-dataset evaluation, we assess the impact of sythetic data on the performance of a method, in the case where the training and the test data of a real-world dataset following a similar data distribution. On the contrary, through the inter-dataset evaluation we assess the impact of synthetic data, in the case where the test set contains positive or/and negative samples that are visually distinct from

the training set. The inter- and intra-dataset performances of the models are reported in Tab V.

TABLE V: Average Precision at 0.5 IoU ($AP_{0.5}$) in all training setups. "R" refers to using only real data for training, while "R + S" refers to using real and synthetic data for training.

Test set (R only)	Training Setups					
	dataset 1		dataset 2		dataset 3	
	R	R+S	R	R+S	R	R+S
Dataset 1	**90.12%**	89.48%	14.49%	**30.68%**	70.84%	**79.55%**
Dataset 2	56.01%	**59.68%**	**81.92%**	71.14%	68.47%	**70.97%**
Dataset 3	41.56%	**53.68%**	25.12%	**31.46%**	**93.51%**	92.90%

In the intra-dataset evaluation, the precision of all models is decreased when synthetic data are inserted in their training sets. This performance drop is minimal ($< 0.1\%$) on two of the three cases. In Dataset 2, the performance drop of the model may be more significant due to the larger synthetic to real-world samples ratio of the training set. On the other hand, the results on the inter-dataset evaluation demonstrated that a model trained on both synthetic and real data and then tested on real data, which are not very similar to those it was trained on, manages to outperform a model which was trained only on real data. The difference in results between the intra-dataset evaluation and the inter-dataset evaluation arises from the fact that the synthetic data alter the distribution of the training set, making the model trained on both the synthetic and the real data, less overfitted on the initial real data and more robust to perform on visually dissimilar data.

IV. Conclusions

Accurate visual UAV detection is crucial to various applications related to surveillance and multi-robot systems. Detectors deployed on real-world environments in this task, must be pre-trained on various datasets, in order to acquire an adequate generalization ability. However, most datasets lack in terms of visual diversity.The proposed synthetic data generation method, which uses as input sets of real-world videos and 3D models of UAVs, can create both challenging and realistic data, suitable for training and evaluating UAV detection methods. The conducted evaluation showed that when detectors are trained on a combination of real-world and synthetic data, exhibit an improved generalization performance, achieving better precision rates when evaluated on real datasets that are visually distinct from the corresponding real training data.

Acknowledgment This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 871449 (OpenDR).

References

[1] "The Multidrone Public Dataset," *Available at: http://multidrone.eu/multidrone-public-dataset/*, 2019.

[2] S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan, "Youtube-8m: A large-scale video classification benchmark," 2016.

[3] M. C. Aksoy, A. S. Orak, H. M. Özkan, and S. B., "Drone dataset: Amateur unmanned air vehicle detection," in *Mendeley Data, V4, doi: 10.17632/zcsj2g2m4c.4*, 2019.

[4] A. Bochkovskiy, C.-Y. Wang, and H. Liao, "YOLOv4: Optimal speed and accuracy of object detection," *ArXiv*, vol. abs/2004.10934, 2020.

[5] A. Carrio, S. Vemprala, A. Ripoll, S. Saripalli, and P. Campoy, "Drone detection using depth maps," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2018, pp. 1034–1037.

[6] A. Coluccia, A. Fascista, A. Schumann, L. Sommer, A. Dimou, D. Zarpalas, M. Méndez, D. de la Iglesia, I. González, J.-P. Mercier, G. Gagné, A. Mitra, and S. Rajashekar, "Drone vs. bird detection: Deep learning algorithms and results from a grand challenge," *Sensors*, vol. 21, no. 8, 2021.

[7] E. Kakaletsis, E. Symeonidis, M. Tzelepi, I. Mademlis, T. A., N. Nikolaidis, and I. Pitas, "Computer vision for autonomous UAV flight safety: An overview and a vision-based safe landing pipeline example," *ACM Computing Surveys*, 2021.

[8] V. Karis and E. Games, "Real shading in Unreal Engine 4," in *SIGGRAPH Courses: Physically Based Shading Theory Practice*, 2013.

[9] D. H. Li, J.and Ye, M. Chung, T.and Kolsch, J. Wachs, and C. Bouman, "Multi-target detection and tracking from a single camera in unmanned aerial vehicles (uavs)," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2016, pp. 4992–4997.

[10] 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.

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

[12] P. Nousi, I. Mademlis, I. Karakostas, A. Tefas, and I. Pitas, "Embedded UAV real-time visual object detection and tracking," in *Proceedings of the IEEE International Conference on Real-time Computing and Robotics (RCAR)*, 2019.

[13] M. Pawełczyk and M. Wojtyra, "Real world object detection dataset for quadcopter unmanned aerial vehicle detection," *IEEE Access*, vol. 8, pp. 174 394–174 409, 2020.

[14] A. Rozantsev, V. Lepetit, and P. Fua, "On rendering synthetic images for training an object detector," *Computer Vision and Image Understanding*, vol. 137, pp. 24–37, 2015.

[15] S. Shah, D. Dey, C. Lovett, and A. Kapoor, "Airsim: High-fidelity visual and physical simulation for autonomous vehicles," in *Field and Service Robotics (FSR)*, 2017.

[16] F. Svanström, C. Englund, and F. Alonso-Fernandez, "Real-time drone detection and tracking with visible, thermal and acoustic sensors," *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 7265–7272, 2021.

[17] 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.

[18] 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.

[19] M. Tzelepi and A. Tefas, "Human crowd detection for drone flight safety using convolutional neural networks," in *Proceedings of the 25th European Signal Processing Conference (EUSIPCO)*, 2017, pp. 743–747.

[20] A. R. Vetrella, R. Opromolla, G. Fasano, D. Accardo, and M. Grassi, "Autonomous flight in GPS-challenging environments exploiting multi-UAV cooperation and vision-aided navigation," in *Proceedings of the AIAA Information Systems-AIAA Infotech @ Aerospace*, 2017.

[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.