Training Deep Learning Models via Synthetic Data: Application in Unmanned Aerial Vehicles

Andreas Kamilaris[1][2], Corjan van den Brink[1] and Savvas Karatsiolis[3]

[1] Pervasive Systems Group, Department of Computer Science
University of Twente, The Netherlands
Email: a.kamilaris@utwente.nl, g.c.vandenbrink@student.utwente.nl
https://www.utwente.nl/en/eemcs/ps/

[2] Research Centre on Interactive Media, Smart Systems and Emerging Technologies (RISE), Nicosia, Cyprus
Email: a.kamilaris@rise.org.cy
http://www.rise.org.cy/

[3] Department of Computer Science, University of Cyprus, Nicosia, Cyprus
Email: skarat01@cs.ucy.ac.cy
https://www.cs.ucy.ac.cy

Abstract. This paper describes preliminary work in the recent promising approach of generating synthetic training data for facilitating the learning procedure of deep learning (DL) models, with a focus on aerial photos produced by unmanned aerial vehicles (UAV). The general concept and methodology are described, and preliminary results are presented, based on a classification problem of fire identification in forests as well as a counting problem of estimating number of houses in urban areas. The proposed technique constitutes a new possibility for the DL community, especially related to UAV-based imagery analysis, with much potential, promising results, and unexplored ground for further research.

Keywords: UAV · Deep Learning · Generative Data · Aerial Imagery

1 Introduction

Deep learning (DL) constitutes a recent, modern technique for image processing and data analysis with large potential [21]. DL belongs to the machine learning (ML) computational field and is similar to artificial neural networks (ANN). DL extends ML by adding more "depth" (complexity) into the model, transforming the data using various functions that allow data representation in a hierarchical way, through several abstraction levels. DL seems to be offering better precision results in classification and/or counting computer vision-related problems, in comparison to traditional techniques such as Scalable Vector Machines and Random Forests, according to relevant surveys [10].

An advantage of DL is the reduced need of feature engineering (FE). Previously, traditional approaches for image classification were based on hand-engineered features, whose performance affected the results heavily [1]. Although

DL does not require FE, it still needs appropriate datasets as input in DL models during learning. These datasets need to be large, to allow DL models to learn the problem elaborately, and expressive, to capture the variation of classes/features that need to be classified/predicted at the model output. An existing problem is the limited availability of such appropriate datasets. This limitation makes DL models sometimes difficult to generalize and to learn the problem well, towards high precision.

Towards addressing this limitation, a recent possibility is the generation of synthetic datasets to train DL models [7], [11], [17]. Models are trained using synthetic images, and they are then able to classify images of the real world, or count objects encountered in the real-world images, via this transfer learning-based method.

The contribution of this paper is twofold: on one hand, to present state of art research in generating synthetic data for training DL models. On the other hand, to present preliminary work on a classification problem of fire identification in forests and a counting problem of estimating number of houses in urban areas, based on two datasets comprised of aerial photos.

2 Related Work

DL is divided in discriminative and generative models [6]. The former is about predictions/classifications, and the latter about synthesis/generation of data similar to the input datasets. The use of generative data to train DL models is promising, with early attempts in agriculture indicating positive outcomes [10].

Table 1 lists related work in the field of generating training data to train DL models. The year of publication for every paper reveals how modern this technique is. Please note that we avoided adding details about performance metrics and evaluation results for each paper, because each author used different metrics and experimented on different real-world datasets for testing. However, the general conclusion in all papers was that the performance according to the metric(s) used, was better than baseline (i.e. datasets not enhanced with synthetic data) or state-of-art related work.

From Table 1, it is evident that related work has not entered yet the domain of UAV-based imagery analysis. The only exception is Meta-Sim [11], which tries to learn a generative model of synthetic scenes automatically, via probabilistic scene grammars, and then it obtains images and their corresponding ground-truth via a graphics engine. Meta-Sim validates this idea addressing the problem of semantic segmentation of simulated aerial views of simple roadways. Beyond this work, to our knowledge, no other work has focused yet on generative data-based approaches for UAV-based imaging-related applications.

Table 1. Related work in generative data for training DL models.

Year	Purpose	Ref.
2007	Simulating fluorescence microscope images of cell populations for automated image cytometry	[12]
2016	Enhancing soil images coming from X-ray tomography, generating roots to help the model identify the roots from the soils	[3]
2016	Simulating top-down images of overlapping plants on soil background, to classify 23 different weed species and maize.	[4]
2016	Generating fully labeled, dynamic, and photo-realistic proxy virtual world, with a focus on objects of interest, e.g. cars.	[5]
2016	Generating synthetic data for semantic segmentation of outdoor scenes, for recognizing aspects such as roads, buildings, cars, people, lights etc.	[19]
2016	Automatically generating realistic synthetic images with pixel-level annotations for semantic segmentation	[20]
2017	Creating synthetic images to predict number of tomatoes in the images.	[18]
2018	Generating synthetic data to identify melanoma skin cancer.	[7]
2018	Synthetic data for 2D bounding box car detection.	[16]
2018	Generating 3D scenes of visually realistic houses, ranging from single-room studios to multi-storied houses, equipped with a diverse set of fully labeled 3D objects, textures and scene layout, for teaching an agent to navigate in an unseen 3D environment.	[25]
2018	Generating scenes for teaching an artificial agent to execute tasks in a simulated household environment.	[17]
2019	a) Generating data for semantic segmentation of aerial views of roadways. b) Simulating urban scenes for object detection in urban car driving.	[11]

3 Methodology

The general methodology followed in this paper is illustrated in Figure 1. More advanced and recent proposals based on this general methodology will be discussed in Section 5. DL models are trained with synthetic data, and then tested with real-world data. The precision/accuracy results are analyzed and compared with the state-of-art related work (if available), and the observations made are given as feedback to the creation process of the synthetic datasets, to become more detailed and complete (e.g. to include some aspects of the real-world data not included originally, but which affect the model's prediction capabilities).

Our approach in synthetic dataset design would be to understand how DL models perform classification, based on the existing real-world datasets (i.e. problem under study for classification or counting). To achieve this, we take advantage of the work in [15], which allows to visualize what happens inside DL models, i.e. which aspects/characteristics of the image are the ones that trigger the final classification. These characteristics could then be used to better design the synthetic datasets, emphasizing on these aspects when creating the simulated images. In this paper, we focused on two different applications:

1. A classification problem of identifying fires in forest areas from aerial photos.

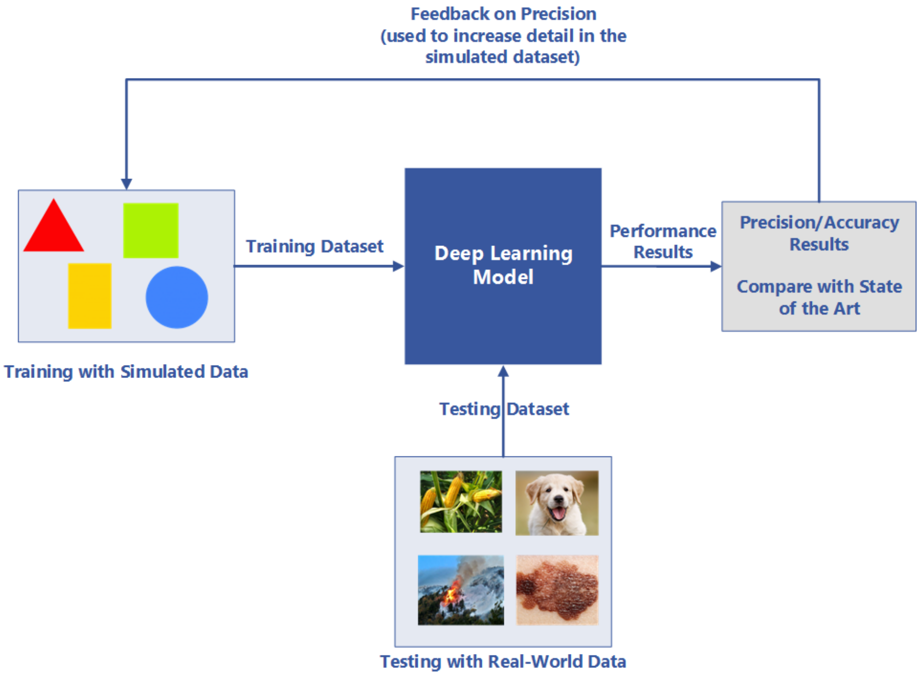
Feedback on Precision
(used to increase detail in the simulated dataset)
Training Dataset
Deep Learning Model
Performance Results
Precision/Accuracy Results
Compare with State of the Art
Training with Simulated Data
Testing Dataset
Testing with Real-World Data

Fig. 1. Basic methodology in generating data for training DL models.

2. A counting problem of estimating number of houses from aerial photos.

The former is useful for UAV which monitor forest areas for fires and smoke, while the latter would be useful for policy-makers who want to understand distribution of houses in urban areas, possibilities for photovoltaic systems, urban gardening in roofs etc.

For the problems under study, the synthetic datasets (used for training the DL model) have been created by means of Python, by using the Python Imaging Library[4] and OpenCV[5]. PIL libraries allow to combine graphics creation, together with programming code and computer logic, using code in order to create dots, lines, rectangles, polygons, circles, ellipses and combinations, allowing to add color, transparency, borders and outlines, but also to include filters such as "Gaussian Blur", smoothen the image, enhance the edges etc. By means of Python scripts, based on the PIL graphic features, we created more complex structures such as smoke, fire, houses, trees, fences, gardens etc. Samples of the synthetic data for the scenarios under study are depicted in Figure 2 (top).

Regarding the real-world datasets (used for testing the DL model), for the fire identification case, 100 aerial photos were downloaded from Google Images, 50 of

[4] Python Imaging Library. https://pypi.python.org/pypi/PIL
[5] OpenCV. https://pypi.python.org/pypi/opencv-python

them showing forest areas and another 50 showing a forest fire. For the counting houses case, 20 aerial photos from urban areas of Tanzania have been selected, from the Open AI Tanzania Challenge[6]. We cropped these photos in 100x100 pixel images, and counted the number of houses manually at each cropped photo. The result was a dataset of 60 images, each having $[0, 38]$ houses from an aerial view. Samples of the real-world datasets for the two scenarios under study are depicted in Figure 2 (bottom). Table 2 describes the number of images used for training and testing of the two scenarios under study.

Table 2. Number of images used for training and testing of the DL models.

Scenario	Purpose	No. of images
Fire identification	Training	2,000 synthetic images
Fire identification	Testing	100 real-world aerial photos (classified as 50 images of forest and 50 images of fire)
Counting houses	Training	10,000 synthetic images (labelled with exact number of houses)
Counting houses	Testing	60 real-world aerial photos (labelled with exact number of houses)

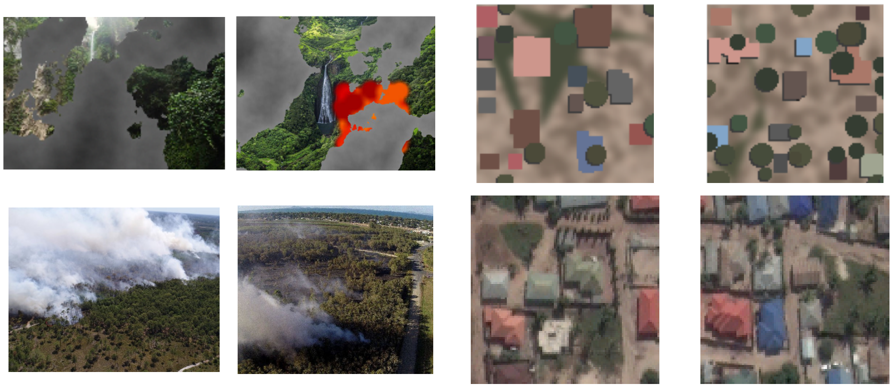

Fig. 2. Example images from the synthetic datasets (top). Example images from the real-world datasets (bottom). Images on the left are for the fire identification scenario, while images on the right for the case of the estimation of number of houses.

As a DL model, we used the Inception-v3 convolutional neural network (CNN) architecture [23] (with some adaptations, see below), as it is one of the fastest CNN architectures available, with high accuracy [2]. We used the default

[6] Open AI Tanzania Challenge. https://blog.werobotics.org/2018/08/06/welcome-to-the-open-ai-tanzania-challenge/

class provided by Keras/TensorFlow during our experiments. Data augmentation was used too. For the counting houses case, our early experiments indicated we should perform adaptations to Inception-v3, to become more optimized for counting correctly. The most important design considerations were the following:

- No pre-training with other datasets (e.g. ImageNet). Filters created by ImageNet are different than the filters required for counting houses.
- Use of dropout (i.e. 35%).
- Max pooling instead of average pooling.
- Use larger filters at the convolutions at the beginning (i.e. 7x7) of the CNN.
- Use a larger value for stride (i.e. stride=5).
- Use a dense layer with only one output at the end of the CNN.
- Use ReLu for the prediction of final outcome.

4 Results

Figures 3 and 4 show the results of the training of the DL models (i.e. synthetic data) and of the testing of the model in real-world data, for the fire identification and counting houses case respectively. Classification accuracy (CA) was used as the performance metric for the fire identification case, while Mean Square Error (MSE) for the counting houses scenario. The fire identification case required 24 epochs of training for the model to learn how to classify with $CA = 96\%$ on the validation dataset, while counting houses needed 18 epochs for the model to learn how to count with $MSE = 20$ on the validation dataset. In this case, MSE measures the average of the squares of the errors of the difference between the actual counts of houses in the images (i.e. ground truth counts of the real-world dataset) and the counts predicted by the DL model.

5 Discussion

Results of the two scenarios under study indicate that synthetic data can prove useful for training DL models, particularly related to UAV-based aerial imagery. This evidence is backed by related work, listed in Section 2. Nevertheless, we need to be cautious with these indications, because the DL models were optimized to perform well in the specific validation datasets. It is questionable (and it has not been tested) whether the DL models can produce similar results in different real-world datasets that focus on similar problems and applications **??**.

The DL model for the fire identification case had very high CA. We achieved this accuracy via a hybrid approach, adding background of real forest images to the generated smoke and fire. Before this, validation CA was around 86%. On the other hand, the DL model for the counting problem still needs improvement. A $MSE = 20$ means that the model can predict the number of houses with an error of ±4.47 houses. For example, for a photo with 20 houses, the model would predict in the range of $[16, 24]$. There is definitely space for further work on this. We note that we reached this MSE after many weeks of observations and the

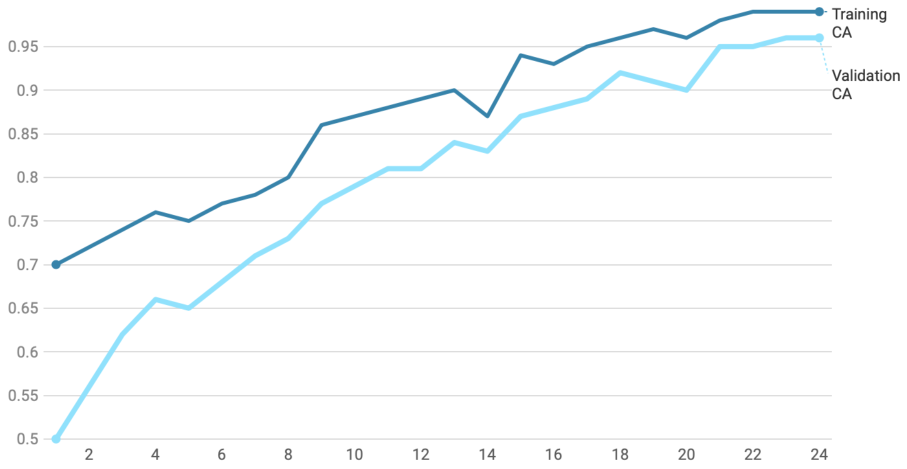

Fig. 3. Training and validation classification accuracy at the fire identification challenge.

iterative process of adding more details to the generated synthetic dataset (see Figure 1). These details included trees, grass, swimming pools, fences etc.. Each of them helped to reduce the overall error.

Applications of the proposed approach can be found in various research domains and scientific disciplines, such as agriculture, life sciences, microbiology, earth sciences etc. The approach of generative data for training DL models would be extremely useful for UAV and robotics [17], [25], where computer vision is involved. It could improve operation and accuracy of automatic robots collecting crops, removing weeds or estimating yields of crops [18], [10]. It could also be used in disaster monitoring and surveillance [9], where remote sensing (i.e. satellites or UAV) is used to identify events of interest (e.g. disasters, violence incidents, land cover mapping, effects on climate change etc.). Finallly, it could be used in environmental studies, e.g. to understand the environmental impact of livestock agriculture [8].

Moreover, we highlight some recent state-of-art work in this domain, which relates to our proposed methodology, showing promising results in application areas other than UAV aerial imagery. First, the work in [24] incorporates two significant improvements: layered boosting (i.e. a layered approach, where training is done in stages) and selective sampling (i.e. streamline the training process by reducing the impact of the low quality samples, such as trivial cases or outliers). Second, the concept of Structured Domain Randomization (SDR) places objects and distractors randomly according to probability distributions [16] and from probabilistic scene grammars [11], which arise from the specific problem at hand. In this manner, SDR-generated imagery enables the neural network to take the context around an object into consideration during detection. Third,

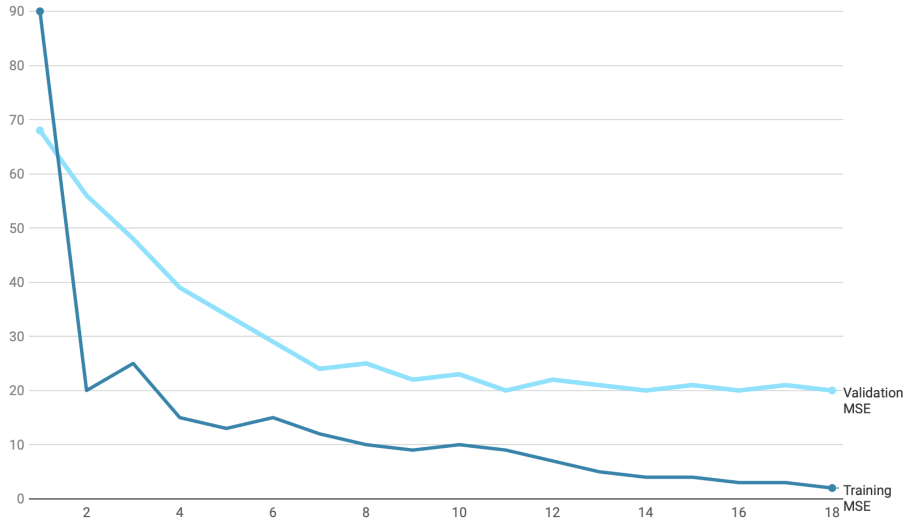

Fig. 4. Training and validation MSE at the counting houses challenge.

related specifically to counting, the work in [13] evades the hard task of learning to detect and localize individual object instances. Instead, it casts the problem as that of estimating an image density whose integral over any image region gives the count of objects within that region. Furthermore, our work, as well as the aforementioned promising approaches [24], [16], [13] can be combined with Generative Adversarial Networks (GANs), to stylize synthetic images to look more like those captured in the real world [14], [26].

Finally, we note another recent possibility, that of utilizing the *Aerial Informatics and Robotics* platform [22] for generating seamlessly training data related to UAV-based aerial imagery. This platform acts as an easy-to-use simulator that aims to enable designers and developers of robotic systems to generate graphical data. Its biggest advantage is that it uses recent advances in computation and graphics to simulate the physics and perception such that the environment realistically reflects the actual world.

6 Conclusion

This paper has described preliminary work in the approach of generating synthetic training data for facilitating the learning procedure of DL models, with a focus on UAV-based aerial imagery. The general methodology of this approach was described, and preliminary results were presented, focused on two different challenges: a classification problem of fire identification in forests as well as a counting problem of estimating number of houses in urban areas. Results were promising, but there is still space for improvement, especially in the counting houses case. Use of synthetic data for training DL models in aerial imagery is a new exciting possibility for the research community working in this area, especially in cases where ground-truth data is scarce or expensive to produce.

For future work, we aim to experiment with more realistic generation of synthetic data, by using game engines such as the Unity development platform[7]. We also aim to apply our methodology in new UAV-related applications such as human crowd counting, identification and counting of endangered species and wild animals etc., enhancing our methodology with the new proposals described in Section 5, in order to minimize the distribution gap between the rendered outputs of the synthetic data and the target real-world data.

References

1. Amara, J., Bouaziz, B., Algergawy, A., et al.: A deep learning-based approach for banana leaf diseases classification. In: BTW (Workshops). pp. 79–88 (2017)
2. Canziani, A., Paszke, A., Culurciello, E.: An analysis of deep neural network models for practical applications. arXiv preprint arXiv:1605.07678 (2016)
3. Douarre, C., Schielein, R., Frindel, C., Gerth, S., Rousseau, D.: Deep learning based root-soil segmentation from x-ray tomography. bioRxiv p. 071662 (2016)
4. Dyrmann, M., Mortensen, A.K., Midtiby, H.S., Jorgensen, R.N., et al.: Pixel-wise classification of weeds and crops in images by using a fully convolutional neural network. In: Proceedings of the International Conference on Agricultural Engineering, Aarhus, Denmark. pp. 26–29 (2016)
5. Gaidon, A., Wang, Q., Cabon, Y., Vig, E.: Virtual worlds as proxy for multi-object tracking analysis. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4340–4349 (2016)
6. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in neural information processing systems. pp. 2672–2680 (2014)
7. Kamilaris, A.: Simulating training data for deep learning models. In: Machine Learning in the Environmental Sciences Workshop, in Proc. of EnviroInfo 2018. Munich, Germany (September 2018)
8. Kamilaris, A., Assumpcio, A., Blasi, A.B., Torrellas, M., Prenafeta-Boldu, F.X.: Estimating the environmental impact of agriculture by means of geospatial and big data analysis: The case of catalonia. In: Proc. of EnviroInfo. Luxembourg (September 2017)
9. Kamilaris, A., Prenafeta-Boldu, F.X.: Disaster monitoring using unmanned aerial vehicles and deep learning. In: Disaster Management for Resilience and Public Safety Workshop, in Proc. of EnviroInfo2017. Luxembourg (September 2017)
10. Kamilaris, A., Prenafeta-Boldu, F.X.: Deep learning in agriculture: A survey. Computers and Electronics in Agriculture **147**, 70–90 (2018)
11. Kar, A., Prakash, A., Liu, M.Y., Cameracci, E., Yuan, J., Rusiniak, M., Acuna, D., Torralba, A., Fidler, S.: Meta-sim: Learning to generate synthetic datasets. arXiv preprint arXiv:1904.11621 (2019)
12. Lehmussola, A., Ruusuvuori, P., Selinummi, J., Huttunen, H., Yli-Harja, O.: Computational framework for simulating fluorescence microscope images with cell populations. IEEE transactions on medical imaging **26**(7), 1010–1016 (2007)

[7] Unity. https://unity.com

13. Lempitsky, V., Zisserman, A.: Learning to count objects in images. In: Advances in neural information processing systems. pp. 1324–1332 (2010)
14. Li, P., Liang, X., Jia, D., Xing, E.P.: Semantic-aware grad-gan for virtual-to-real urban scene adaption. arXiv preprint arXiv:1801.01726 (2018)
15. Olah, C., Satyanarayan, A., Johnson, I., Carter, S., Schubert, L., Ye, K., Mordvintsev, A.: The building blocks of interpretability. Distill **3**(3), e10 (2018)
16. Prakash, A., Boochoon, S., Brophy, M., Acuna, D., Cameracci, E., State, G., Shapira, O., Birchfield, S.: Structured domain randomization: Bridging the reality gap by context-aware synthetic data. arXiv preprint arXiv:1810.10093 (2018)
17. Puig, X., Ra, K., Boben, M., Li, J., Wang, T., Fidler, S., Torralba, A.: Virtualhome: Simulating household activities via programs. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8494–8502 (2018)
18. Rahnemoonfar, M., Sheppard, C.: Deep count: fruit counting based on deep simulated learning. Sensors **17**(4), 905 (2017)
19. Richter, S.R., Vineet, V., Roth, S., Koltun, V.: Playing for data: Ground truth from computer games. In: European Conference on Computer Vision. pp. 102–118. Springer (2016)
20. Ros, G., Sellart, L., Materzynska, J., Vazquez, D., Lopez, A.M.: The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3234–3243 (2016)
21. Schmidhuber, J.: Deep learning in neural networks: An overview. Neural networks **61**, 85–117 (2015)
22. Shah, S., Dey, D., Lovett, C., Kapoor, A.: Aerial informatics and robotics platform. Washigton: Microsoft Research (2017)
23. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2818–2826 (2016)
24. Walach, E., Wolf, L.: Learning to count with cnn boosting. In: European conference on computer vision. pp. 660–676. Springer (2016)
25. Wu, Y., Wu, Y., Gkioxari, G., Tian, Y.: Building generalizable agents with a realistic and rich 3d environment. arXiv preprint arXiv:1801.02209 (2018)
26. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision. pp. 2223–2232 (2017)