Synthetic data for automatic target recognition from small drones

Celso de Melo, Damon Conover, Domenick Poster, Sarah Leung, Robert Nguyen, Joseph Conroy
DEVCOM Army Research Laboratory, 2800 Powder Mill Rd, Adelphi, MD 20783

ABSTRACT

Automatic target recognition (ATR) technology is likely to play an increasingly prevalent role in maintaining situational awareness in the modern battlefield. Progress in deep learning has enabled considerable progress in the development of ATR algorithms; however, these algorithms require large amounts of high-quality annotated data to train and that is often the main bottleneck. Synthetic data offers a potential solution to this problem, especially given recent proliferation of tools and techniques to synthesize custom data. Here, we focus on ATR, in the visible domain, from the perspective of a small drone, which represents a domain of growing importance to the Army. We describe custom simulators built to support synthetic data for multiple targets in a variety of environments. We describe a field experiment where we compared a baseline (YOLOv5) model, trained on off-the-shelf large generic public datasets, with a model augmented with specialized synthetic data. We deployed the models on a VOXL platform in a small drone. Our results showed a considerable boost in performance when using synthetic data of over 40% in target detection accuracy (average precision with at least 50% overlap). We discuss the value of synthetic data for this domain, the opportunities it creates, but also the novel challenges it introduces.

Keywords: Automatic Target Recognition, Synthetic Data, Small Drones

1. INTRODUCTION

Automatic target recognition (ATR) is a key capability to enhance situational awareness in modern, complex, dynamic battlefields. ATR consists of detection of objects and targets in sensor data on manned and unmanned air and ground platforms [1]. Object detection is a well-known problem in the scientific computer vision community that seeks to detect an object of interest in imagery or video – the last decade has seen an explosion of deep learning approaches that have achieved commercial success in applications ranging from face recognition to autonomous driving [2]. A natural extension of the problem pertains to the automatic recognition of activity and threats. Just like for object detection, deep learning techniques have been increasingly successful in activity recognition tasks [3]. However, these deep learning models rely on large quantities of labeled data for training, which is often the main bottleneck, especially for military use cases [4, 5]. Consequently, deep learning models trained on large scale public datasets such as ImageNet are not performant on operational data and cannot be readily retrained due to the limited availability of real labeled data collected in operationally relevant environments. The development of AI/ML algorithms capable of accurately detecting objects and activity is therefore essential for ATR applications, though constrained by the acute lack of extensive labeled training data.

2. WHY SYNTHETIC DATA?

To demonstrate the value of synthetic data, we describe a field experiment conducted where the purpose was for a small drone to detect and track a white truck. The target was intentionally chosen to be a civilian vehicle to allow for comparison of a baseline model trained on a public dataset versus a model fine-tuned on synthetic data. The baseline model was pretrained on the COCO dataset [6], which is a large-scale object detection dataset with 80 object categories, including trucks. To generate synthetic data, we created a custom Unity simulator to generate imagery of trucks in a variety of poses and with varying parameters for the camera position and illumination conditions

(Figure 1-A). We generated a few thousand synthetic images. The baseline model consisted of a YoloV5 small model pretrained on the COCO dataset. YoloV5 was chosen as it is an object detection algorithm that produces good accuracy while maintaining real-time performance, thus making it an ideal candidate to deploy in low compute platforms. The synthetic model consisted of the baseline model, but fine-tuned on the synthetic data samples. The models were deployed on small drones equipped with the VOXL platform. A field experiment was conducted where a truck would drive on a dirt road and the drone was tasked with detecting and tracking it (Figure 1-B). The results, shown in Figure 1-C, indicate that the synthetic data substantially outperformed the baseline model. This experiment, thus, presents a clear example of the value of synthetic data for improving ATR performance, when compared to off-the-shelf solutions based on generic public data.

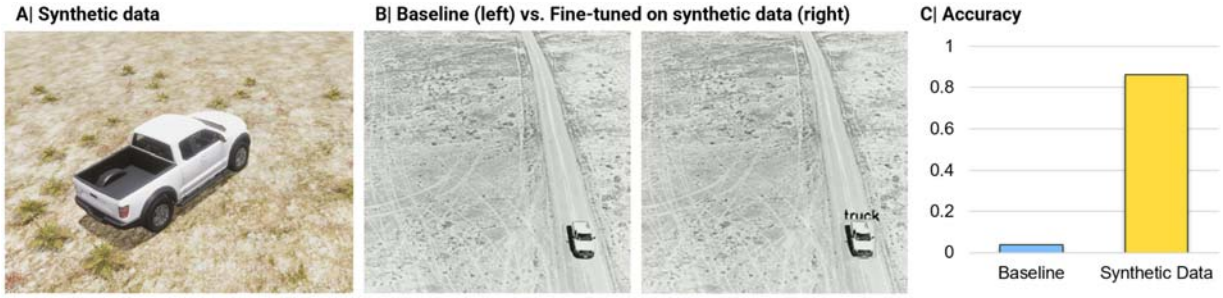

Figure 1. The ATR model fine-tuned on synthetic data outperformed a baseline model trained on off-the-shelf real data. A | Synthetic data imagery generated using our simulator. B | The synthetic data model deployed on a small drone clearly outperformed the baseline model. C | Target recognition accuracy.

3. OPTIMIZING THE VALUE OF SYNTHETIC DATA THROUGH DOMAIN RANDOMIZATION

Domain randomization consists of varying the parameters used to generate synthetic data, so that the dataset broadly captures the distribution in the target domain. By training on such a diverse dataset, the hope is that the model will be more robust to variation in the target domain and generalize better to novel samples. To illustrate this technique, we describe another experiment where the goal was to detect military targets. Once again, we created a Unity simulator that was able to generate imagery for these targets in a variety of poses, backgrounds, illumination, and perspectives (Figure 2-A). A large synthetic dataset, consisting of several thousand samples, was created by systematically varying these parameters, such that each target would be rendered in the same diverse conditions. We trained a YoloV5 model on this synthetic data. Notice that, in contrast to the prior experiment, there was no pretraining on real data. To test this model, we relied on a real dataset collected for these same targets, taken from a small drone perspective. The real dataset was split into training, evaluation and test sets. For our baseline, we trained a YoloV5 model on real data and tested on the test set. As shown in Figure 2-B in the blue bars, this baseline achieved almost perfect performance in this particular dataset – which was likely due to the proximity of the targets and absence of clutter. However, more importantly, we note that the performance of the model trained only on synthetic data (Figure 2-B, red bars) was well above chance, emphasizing again the value of synthetic data. Nevertheless, this experiment also illustrates an important challenge that occurs when training on

synthetic data and deploying/testing with real data: there is often a *domain gap* in the performance when shifting from the synthetic to the real domain. We focus on this issue in Section 5.

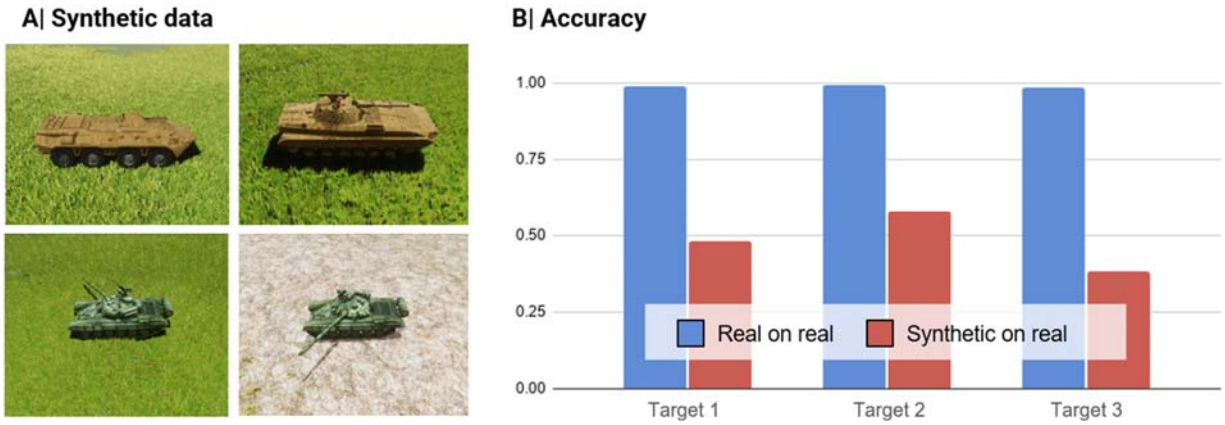

Figure 2. Training only on synthetic data can lead to reasonable performance, but there is often a synthetic-to-real domain gap in performance. A | Diverse synthetic data generated with our simulator. B | The model trained only on synthetic data led to above-chance performance, but there was still a significant gap vis-à-vis the model trained on real data.

4. BRINGING THE BEST OF BOTH WORLDS

In this section we illustrate another method of leveraging synthetic data, which consists of *mixing* real and synthetic data. The idea is that mixing data allows different data types to strengthen training where others may have weaknesses (e.g., synthetic data tends to be more diverse, but real data may capture low level details better). We illustrate this idea in the task of recognizing human activity using pure-RGB sensors – i.e., no depth information is used. This is a domain of increasing practical relevance, given the need to monitor the surrounding environment for relevant activity, such as threats. In this case, we looked at control gestures – e.g., halt, follow me – which can be used to control ground or aerial unmanned systems. We collected samples from a few subjects and split this dataset into training, evaluation, and test sets [3]. We then created another Unity simulator for these gestures, which was able to generate diversity in the environment (e.g., background), character (e.g., body shape), and gesture performance (e.g., speed). For this experiment, we considered the I3D model, which is a state-of-the-art activity recognition model. We trained three models: one trained only on real data, another trained only on synthetic data, and a final one trained on mixed data (50% of each). The results, shown in Figure 3-B, replicate the typical domain gap in performance between the real and the synthetic models. However, the interesting result was that mixing the two types of data led to substantially better performance than the model trained only on real data.

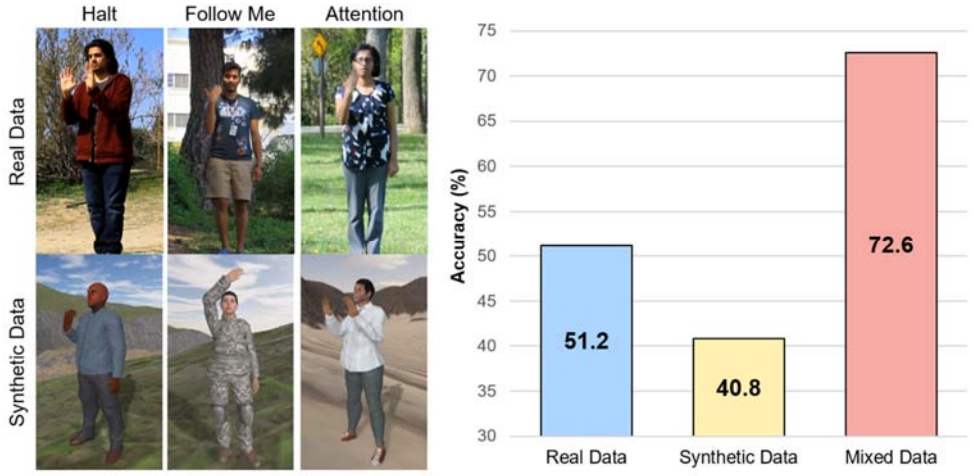

Figure 3. Mixing real and synthetic data led to the best action recognition performance. On the left, samples of the real and synthetic videos are shown. On the right, performance for models trained on real data vs. synthetic data vs. mixed data is shown.

5. CLOSING THE DOMAIN GAP

In addition to the synthetic-to-real domain shift, another common problem when considering drone applications is the ground-to-air domain shift. This happens because often it is easier to obtain ground data than aerial data; however, there are important differences (e.g., geometry of the motion) between the ground and aerial perspectives. Therefore, in many applications, we need methods that support generalization across a variety of domain shifts, in particular, the synthetic-to-real and ground-to-air shifts. To support the development and comparison of domain adaptation methods we worked on developing a dataset that has various control gesture performances, from the ground and air perspectives, and using real and synthetic data (Figure 4-A) [7]. To be precise, here we focus on the *unsupervised domain adaptation* problem, where we have abundant source labeled data (e.g., synthetic data) and a small number of target unlabeled data (e.g., real data) [8]. The goal is to develop an algorithm that is trained on the source and target data and is able to be deployed in the target domain. Several algorithms have begun being proposed to tackle this challenge and, here, we focus on CO^2A, which is a recent unsupervised domain adaptation state-of-the-art algorithm. We ran various experiments, as shown in Figure 4-B, focusing on the synthetic-to-real shift for the ground and air perspectives. The results show that the performance when training on real data only outperforms the model trained on synthetic data only, for both perspectives. This is a replication of the typical synthetic-to-real domain gap. However, when using the domain adaptation method, we see improvement in the performance for both perspectives, albeit still not as good as the model trained on real data only, which suggests room for future improvement. Moreover, we note that the performance for the models trained on the ground data tend to perform better than the models trained in the air data, thus emphasizing the different characteristics between these perspectives. We recently began exploring a method that relies on intermediate 3D representations to support robust generalization from the ground to the air perspective [9].

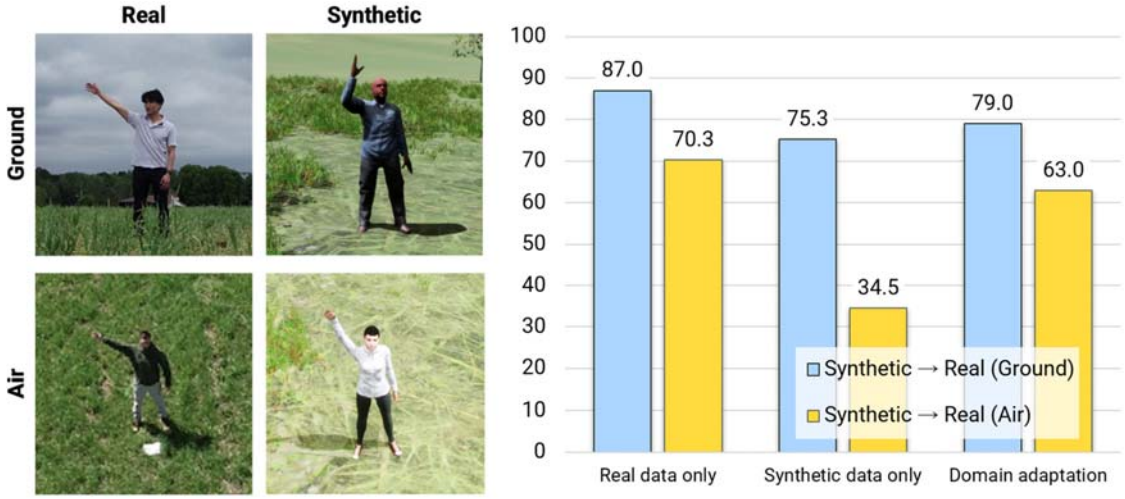

Figure 4. Domain adaptation methods can help mitigate the performance impact due to the synthetic-to-real and ground-to-air domain shifts.

6. CONCLUSION

Synthetic data is likely to play an increasingly important role in the development of ATR algorithms, including for small drones applications. There is growing evidence that synthetic data can mitigate the need for large quantities of labeled real data, as also shown in our experiments here. Moreover, the proliferation of tools to synthesize data, such as game engines and generative models, reduces the barriers to create synthetic data. Further work is still needed, nevertheless, to tackle the synthetic-to-real shift, but current trends in domain adaptation algorithms show much promise. Finally, by integrating the synthesis and machine learning pipelines it becomes possible to optimize the synthesis generation process and support novel learning paradigms, such as continual learning.

REFERENCES

1. Kott, A., and Alberts, D. 2017. How do you command an Army of intelligent things? *Computer*, 50(12).
2. Liu, L., Ouyang, W., Wang, X., Fieguth, P., Chen, J., Liu, X., and Pietikäinen, M. 2020. Deep learning for generic object detection: a survey. *International Journal of Computer Vision volume*, 128.
3. de Melo, C., Rothrock, B., Gurram, P., Ulutan, O., & Manjunath, B. S. (2020) Vision-based gesture recognition in human-robot teams using synthetic data. In *Proceedings of the International Conference on Intelligent Robots and Systems (IROS)*.
4. Rao, R., de Melo, C., & Krim, H. (2021) Synthetic environments for artificial intelligence (AI) and machine learning (ML) in multi-domain operations. *DEVCOM Army Research Laboratory*, May, ARL-TR-9198.
5. de Melo, C., Torralba, A., Guibas, L., DiCarlo, J., Chellappa, R., & Hodgins, J. 2021. Next-generation deep learning based on simulators and synthetic data. *Trends in Cognitive Sciences*, 26(2).
6. Lin, T., et al. (2014) Microsoft COCO: Common Objects in Context. In *Proceedings of European Conference on Computer Vision (ECCV)*.
7. Reddy, A., Shah, K., Paul, W., Mocharla, R., Hoffman, J., Katyal, K., Manocha, D., de Melo, C., & Chellappa, R. (2023) Synthetic-to-real domain adaptation for action recognition: A dataset and baseline performances. In *Proceedings of International Conference on Robotics and Automation (ICRA)*.
8. Oza, P., Sindagi, V., Vibashan, V., and Patel, V. (2021). Unsupervised domain adaptation of object detectors: A survey. *ArXiv*, arXiv:2105.13502.
9. Shah, K., Shah, A., Lau, C., de Melo, C., & Chellappa, R. (2023) Multi-view action recognition using contrastive learning. In *Proceedings of Winter Conference on Applications of Computer Vision (WACV)*.