Adversarial Learning for Effective Detector Training via Synthetic Data*

Vadim Gorbachev[0000−0002−4929−0083],
Andrey Nikitin[0000−0002−9810−9965], and Ilya Basharov[0000−0002−6442−7789]

FSUE «GosNIIAS», Moscow, Russia
{vadim.gorbachev, nad, basharov.iv}@gosniias.ru

Abstract. Current neural network-based algorithms for object detection require a huge amount of training data. Creation and annotation of specific datasets for real-life applications require significant human and time resources that are not always available. This issue substantially prevents the successful deployment of AI algorithms in industrial tasks. One possible solutions is a synthesis of train images by rendering 3D models of target objects, which allows effortless automatic annotation. However, direct use of synthetic training datasets does not usually result in an increase of the algorithms' quality on test data due to differences in data domains. In this paper, we propose the adversarial architecture and training method for a CNN-based detector, which allows the effective use of synthesized images in case of a lack of labeled real-world data. The method was successfully tested on real data and applied for the development of unmanned aerial vehicle (UAV) detection and localization system.

Keywords: Detection, Domain Adaptation, Neural Networks, UAV, Adversarial Training, Synthetic Data

1 Introduction

Object detection is one of the key tasks of computer vision. The main purpose of detection is to locate, identify, and localize all objects of certain classes in an image. The well-established approach is using convolutional neural networks of various architectures (CNNs). Currently, a number of detectors have achieved the ability to work in real-time with fairly high accuracy.

While excellent performance has been achieved on large public datasets, real-world object detection still faces great difficulties. One reason for this is the lack of sufficient annotated real-world data to train detection algorithm for a specific task. Another reason is the difference between filming conditions at the training and the execution stages. View angles peculiarities, object appearance, background, illumination, image quality determine as so-called "domain" of data. The development of domain-invariant recognition methods or image domain transfer methods is a complicated scientific task. Manual target dataset creation and annotation is an expensive and time-consuming problem,

* Publication is supported by RFBR grant 19-07-00844

which researchers try to bypass in different ways. In particular, the design of a domain-invariant object detection algorithm is an important and prospective task.

In this paper, the domain adaptation problem is investigated in relation to the object detection task for mini-UAVs (drones). The source domain is the data obtained by rendering 3D models of the object (artificial data), the target domain is a limited number of real-world images. Our main achievement is the application of domain transfer by adversarial training technique between synthetic and real data, which allowed us to achieve high detection accuracy provided an extremely limited amount of real data. Such results are not achievable by conventional training methods as we demonstrated in experiments.

The drone detection algorithm is a key component for the development of a passive indoor drone positioning system (where there is no signal from satellite navigation systems available) using a set of stationary surveillance cameras [1]. Compared to our previous work [2], this article describes the more complicated task of detection, which required the development of new, more precise, and complex methods. The increased task complexity was caused, firstly, by longer distance to the object, which led to a lower target object size on images (also in relation to frame size) of the new camera set. Secondly, we had to detect the DJI Mavic 2 Pro drone, which is of smaller size and less contrast to the background.

2 Related Works

The idea of automatic object detection on images has been around for a long time. The first successful attempts were based on detection of low-level image features, for example, the Canny edge detector [3] or correlation algorithms for comparison between objects and a template. Modern image analysis methods based on neural networks have significantly outperformed classical algorithms in terms of accuracy. Neural network architectures for detection are divided into three main types:

1. One-stage methods – You Only Look Once (YOLO) [4,5,6,7], Single-shot Multibox Detector (SSD) [8,9] and etc. The main idea is that the image is divided into regions and features are extracted in each one. Then the network predicts bounding boxes and probabilities for each region. Repeated candidates are discarded by the non-maximum suppression algorithm. The methods are characterized by a high FPS.
2. Two-stage methods – region-based CNN algorithms and analogs [10,11,12]. In the first stage features extracted from an image are fed into Region Proposal Network, and in the second stage a class is predicted and the bounding boxes are additionally regressed in candidate regions (after their alignment). The algorithms achieve higher accuracy than single-stage algorithms but have slightly lower performance.
3. Cascaded methods – Cascade R-CNN [13] and others. There are characterized by learning a sequence of detectors with increasing thresholds. More complex cascade architecture that adapts errors from other levels helps to boost quality significantly, but a model contains more parameters and becomes computationally more sophisticated.

Domain adaptive methods have been developed last years. They have made it possible to expand the neural network algorithm's applicability and partially solve the lack of annotated data problems for applied tasks.

The domain adaptation concept based on adversarial training was proposed in [14] to reduce the difference between semantically identical neural network representation but visually disparate data from different domains. A network was trained to solve two problems simultaneously: the target task is the classifying objects, and the side task is the classifying the domain. In this case, the first problem solution was stimulated, and the side problem was penalized. Due to this, the data representation invariance was achieved. This approach was successfully applied to a wide range of other problems.

The A-Fast-RCNN detector [15] was able to detect occlusion and deformation by training on images generated by the GAN. Domain Adaptive Faster R-CNN [16] improved the detection object quality on different types of images by applying an adversarial domain adaptation to both levels the image and the instance. The paper [17] deals with detection invariant to view angle, object scale, and weather conditions by means of adversarial training. Siamese-GAN [18] is suitable for analyzing invariant features for both annotated and non-annotated images coming from two different domains. CyCADA [19] is a unified cyclic-serial network with an adversarial loss function to provide the domain invariance. Also are known works are devoted to detector training on data from alternative sources, such as images obtained by 3D models [20,21] or from artificially generated data [22]. In this work, we present adversarial training method and architecture for object detection.The proposed method is compared with other existing detection methods (fine-tune, augmentation, etc.) and shows better results.

3 Data Preparation

Fig. 1. 3D drone model render.

Neural networks require lot of annotated training samples to achieve outstanding accuracy. Those algorithms have large number of tunable parameters that determines their

high flexibility. Therefore, a large number of annotated images are required to train these CNN-based algorithms. In case of detection, annotation indicates the object class and its coordinates (the bounding box) on the image. To avoid time- and resource-consuming manual data labelling, artificial creation of training and testing images was applied. Images were created by rendering the 3D model of target objects.

During the data creation process, in the 3D modeling system the drone model was drawn on a homogeneous background. Additionally, the object mask was drawn. Then the image and mask were randomly transformed by rotation, scaling, shifting color channels, reflection, perspective transformation, blurring, adding salt and pepper noise. After that, the object image by its mask was placed on background images. Large collection of arbitrary images was used for backgrounds. Local smoothing with a Gaussian kernel were performed at the object borders in order to make such pasting look natural. Distracting objects from the Coil 100 dataset [23] were added to each image to increase the discriminating ability of a detection algorithm.

(a) (b)

Fig. 2. Generated image examples with a random background.

The drone was flown and surveyed from 6 high-resolution video cameras in the test hangar to prepare the test samples and expand the training samples (Fig. 3). Data from all 6 cameras were annotated by hand.

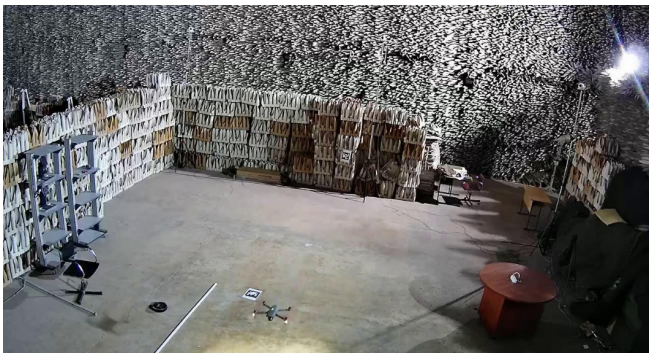

Fig. 3. Example of real target image. The drone at the bottom center is in a test hangar.

4 Adversarial detector training method

Faster R-CNN architecture was used as the base detection algorithm [12] because it shows a high-level accuracy and has efficient implementations capable of working in real-time. The input of the algorithm takes a three-channel image. This image is fed into a CNN, which outputs a feature map. This image is also fed into RPN, where we get ROIs, which may contain the object being detected. Then the ROI Pooling layer converts the feature vector of ROI into a fixed-length feature vector. At the final stage, the bounding boxes are regressed and objects contained in regions of interest are classified.

The problem of most algorithms is underperformance on real data, while training on artificial images. The unlimited synthetic data was available because of 3D rendering. The updated algorithm main goal is training on synthetic data effectively, and then show the good performance on real data. Object artifacts in synthetic images are perceived by the detection algorithm as highly informative features. The domain adaptation method was applied using adversarial training, following the approaches [14,17] to overcome this problem. Feature vectors encoded by the neural network in the object area contain information that most relevant for minimizing the loss function on the training sample, which is determined by the principle of backpropagation. It is clear that there is no guarantee that this representation will contain high-level invariant object features, rather than low-level object features on a particular sample of images (in particular, artifacts of pasting renders of 3D model on images). This is what seems to make it difficult to apply an algorithm trained on synthetic images to real data.

The following approach was applied in order to train the network to create more invariant object representations. Special domain classification branch (head) was added after backbone, in parallel with detection branch. The network was trained to perform object detection and data domain classification simultaneously (synthetic or real-world) based on the attributes computed by the backbone. The new domain classification head was used for network to “forget” the domain. The correct domain classification was penalized by gradient propagation of the respective loss function with the reverse sign. In other words, the network learned to create object representations invariant to the image domain. The modified detection algorithm architecture (Faster R-CNN+AT) is presented in the Figure 4.

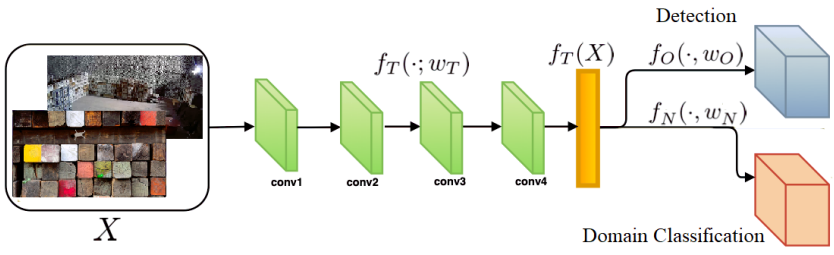

Fig. 4. Proposed Faster R-CNN+AT architecture. The domain classification branch predicts image domain: synthetic or real-world image.

The main role in training of the proposed network is a special loss function. It provides a mathematical formulation to compare predicted and ground-truth and significantly effects on training time and reaching a required accuracy. The minimizing loss function is set as follows:

$$\min_{f_O, f_T} L_O(f_O(f_T(Y_T)), Y_O) - \gamma L_N(f_N(f_T(Y_T)), Y_N), \\ \min_{f_N} L_N(f_N(f_T(X)), Y_N), \tag{1}$$

where O - object detection task, N – domain classification task, T – feature extraction, $f_{O,N,T}$ – task modules, respectively, $Y_{O,N}$ – training samples for the task, respectively, $L_{O,N}$ – losses functions, respectively, γ – adversarial coefficient.

$$l_{1,\text{smooth}}(x) = \begin{cases} |x|, & \text{if } |x| > a \\ \frac{x^2}{|a|} & \text{if } |x| < a \end{cases} \tag{2}$$

Smoothed l_1 (2) is taken as a loss function for the detection task. Often softmax loss is used in classification tasks, but it causes gradient explosions in adversarial training according to [15]. It is replaced by negative entropy function. This helps the model to make "uncertain" predictions about the image domain.

So, the final loss function looks like:

$$\min_{f_O, f_T} L_O(f_O(f_T(Y_T)), Y_O) + \gamma L_{ne}(f_N(f_T(Y_T))), \\ \min_{f_N} L_N(f_N(f_T(X)), Y_N) \tag{3}$$

Training this model is similar to training GANs [24]. The algorithm is shown in Algo. 1.

Algorithm 1 Faster R-CNN+AT with adversarial training for a detection task.

INPUT: pre-trained f_T, f_O, f_N
k - iterations for dumping f_N weights.

procedure TRAINING FASTER R-CNN+AT(f_T, f_O, f_N, k)
  for epoch in range(epochs) **do**
    Sample a mini-batch of n examples $[X_1, \cdots, X_n]$
    Update $f_T(w_T)$ and $f_O(w_O)$ with gradients:
    $\nabla_{w_T, w_O} \frac{1}{n} \sum_{i=0}^{n} L_O(f_O(f_T(Y_T^i)), Y_O) + \gamma L_{ne}(f_N(f_T(Y_T^i)))$
    while predictions from f_N have training accuracy $\leqslant 0.9$ **do**
      Update $f_N(w_N)$ with gradients:
      $\nabla_{w_N} \frac{1}{n} L_{ne}(f_N(f_T(Y_T^i)))$
    end while
    Restart f_N for every k iterations and repeat the Procedure again.
  end for
end procedure

5 Results

The accuracy of the proposed algorithm (Faster R-CNN+AT) has been tested on different datasets to demonstrate its advantage over the base Faster R-CNN, provided different ratios of synthetic and real images. Synthetic data were generated by rendering and automatically labeled. The real data were annotated manually. Data from 5 cameras (800 images) and synthetic data was involved in the training phase. Data from the 6th camera (124 images) was used for testing, which did not participate in the training phase. Obtained accuracy (precision and recall) at fixed cutoff thresholds by probability, as well as standard average metrics (average precision and mean average precision) in Tab. 1.

Two main conclusions can be inferred from the experiments results (Tab. 1).

1. The results demonstrate that the proposed method provides the most effective way (among other approaches) to use synthetic data along with real data than the amount of real data is fixed. Although the use of pure synthetic data is inefficient (first row in Tab. 1), the amount of real data is insufficient to train detector solely on real images (second row in Tab. 1), adversarial training achieves perfect results (last row).
2. The developed method outperforms the base one due to the adversarial learning application. The advantage is observed regardless of what real-world and synthetic images ratio was used. It is determined not only by the complexity of the neural model (the feature extraction network is identical) but by the learning principle.
3. Experiments prove that the proposed method is more effective than the widespread fine-tuning technique, the results of which are shown in the penultimate row of table 1. Fine-tuning is training the model first exclusively on a large available collection of relatively relevant data (in our case, synthetic), then further training on target data (in our case, real). This allows one to get better results than simple data mixing, but still less accurate than the results of the proposed Faster R-CNN+AT method.

6 Conclusion

The CNN-based object detection training algorithm and architecture based on the adversarial technique is proposed. By its application, we solved the problem of the lack of annotated target training data. The training algorithm enforces the detector's encoder subnet to generate domain-invariant image features. The proposed algorithm has the ability to train detectors mainly on synthetic images (obtained by 3D objects rendering) and a limited number of real-world data but shows high accuracy on target real images. We have shown the superiority of the proposed training scheme for training object detectors in a set of experiments using different real and synthesized image ratios in the training set. The method could be applied to arbitrary detector architectures.

Table 1. Accuracy comparison between the proposed detection model (Faster R-CNN+AT) and the base model (Faster R-CNN) at real and synthetic images combined in different ratios at training set.

Adversarial Training	Synthetic images	Real-world images	Precision	Recall	AP@0,5	AP@0,75	mAP
✗	2400	0	0.0286	0.2054	0.0093	0	0.0025
✗	0	800	0.2430	1	0.7222	0.3400	0.3544
✗	800	800	0.4299	0.8214	0.6781	0.3359	0.3455
✓	800	800	0.9912	1	0.9998	0.9530	0.8365
✗	1600	800	0.5077	0.900	0.8272	0.4486	0.4261
✓	1600	800	1	1	1	0.9703	0.8341
✗	2400	800	0.4633	0.9099	0.8224	0.4526	0.4412
✗[1]	2400	800	0.7092	0.8929	0.7947	0.4651	0.4471
✓	2400	800	1	1	1	0.9633	0.8178

[1] Fine-tuning on real-world data based on weights of network trained on synthetic data.

References

1. Blokhinov, Yu., Gorbachev, V., Nikitin, A., Skryabin, S.: Technology for the Visual Inspection of Aircraft Surfaces Using Programmable Unmanned Aerial Vehicles. Journal of Computer and Systems Sciences International, 960–968 (2019). https://doi.org/10.1134/S1064230719060042
2. Gorbachev, V., Blokhinov, Yu., Nikitin, A., Andrienko, E.: Technology for Indoor Drone Positioning Based on CNN Detector. In: Proceedings of the 29th International Conference on Computer Graphics and Vision, pp. 280–284, Bryansk, Russia (2019). https://doi.org/10.30987/graphicon-2019-2-280-284
3. Canny, J.: A computational approach to edge detection. Journal of Pattern Analysis and Machine Intelligence, IEEE Transactions on, 679–698 (1986). https://doi.org/10.1109/TPAMI.1986.4767851
4. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 779–788, IEEE, Las Vegas, NV, USA (2016). https://doi.org/10.1109/CVPR.2016.91
5. Redmon, J., Farhadi, A.: Yolo9000: better, faster, stronger. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7263—7271, IEEE, Honolulu, HI, USA (2017).https://doi.org/10.1109/CVPR.2017.690
6. Redmon, J., Farhadi, A.: Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767 (2018).
7. Bochkovskiy, A., Wang, C.-Y., Liao, H.-Y.: Yolov4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934 (2020).
8. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., Berg, A.: SSD: Single Shot Multibox Detector. In: ECCV 2016, pp. 21—37, Springer, Cham (2016). https://doi.org/10/gc7rk8
9. Fu, C.-Y., Liu, W., Ranga, A., Tyagi, A., Berg, A.: DSSD: Deconvolutional single shot detector. arXiv preprint arXiv:1701.06659 (2017).
10. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 580–587, IEEE, Columbus, OH, USA (2014). https://doi.org/10.1109/CVPR.2014.81

11. Girshick, R.: Fast R-CNN. In: Proceedings of the IEEE international conference on computer vision, pp. 1440—1448, IEEE, Santiago, Chile (2015). https://doi.org/10.1109/ICCV.2015.169
12. Ren, S., He, K., Girshick R., Sun. J.: Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In: IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 1137–1149, IEEE (2016). https://doi.org/10.1109/TPAMI.2016.2577031
13. Cai, Z., Vasconcelos, N.: Cascade R-CNN: Delving into high quality object detection. Journal of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, p. 6154—6162, (2018). https://doi.org/10.1109/CVPR.2018.00644
14. Ganin, Y., Lempitsky V.: Unsupervised domain adaptation by backpropagation. arXiv preprint arXiv:1409.7495 (2014).
15. Wang, X., Shrivastava, A., Mulam, H.: A-Fast-R-CNN: Hard positive generation via adversary for object detection. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3039–3048, IEEE, Honolulu, HI, USA (2017). https://doi.org/10.1109/CVPR.2017.324
16. Chen, Y., Li. W., Sakaridis, C., Dai, D., Van Gool, L.: Domain Adaptive Faster R-CNN for Object Detection in the Wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern, pp. 3339–3348, IEEE, Salt Lake City, UT, USA (2018). https://doi.org/10.1109/CVPR.2018.00352
17. Wu, Z., Suresh, K., Narayanan, P., Xu, H., Kwon, H., Wang, Z.: Delving Into Robust Object Detection From Unmanned Aerial Vehicles: A Deep Nuisance Disentanglement Approach. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 1201–1210, IEEE, Seoul, Korea (South) (2019). https://doi.org/10.1109/ICCV.2019.00129
18. Bashmal, L., Bazi, Y., Alhichri, H., Alrahhal, M., Ammour, N., Alajlan, N.: Siamese-GAN: Learning invariant representations for aerial vehicle image categorization. In: Remote Sensing, pp. 351 (2018). https://doi.org/10.3390/rs10020351
19. Hoffman, J., Tzeng, E., Park, T., Zhu, J.-Y., Isola, P., Saenko, K., Efros, A., Darrell, T.: CyCADA: Cycle-consistent adversarial domain adaptation. arXiv preprint arXiv:1711.03213 (2017).
20. Peng, X., Sun, B., Ali, K., Saenko, K.: Learning deep object detectors from 3D models. In: IEEE International Conference on Computer Vision (ICCV), pp. 1278–1286, IEEE, Santiago, Chile (2015). https://doi.org/10.1109/ICCV.2015.151
21. Sun, B., Saenko, K.: From virtual to reality: Fast adaptation of virtual object detectors to real domains. In: BMVA Press (2014). https://doi.org/10.5244/C.28.82
22. Hattori, H., Naresh Boddeti, V., Kitani K. M., Kanade T.: Learning scene-specific pedestrian detectors without real data. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3819–3827, IEEE, Boston, MA, USA (2015). https://doi.org/10.1109/CVPR.2015.7299006
23. Nene, S. A., Nayar, S. K., Murase, H.: Columbia Object Image Library (COIL-100). Technical Report CUCS-006-96 (1996).
24. 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, Curran Associates, Inc. (2014).