2D Image Object Detection Aided by Generative Adversarial Networks: A Literature Review

Caio Vinicius Bertolini[1*], Roberto Monteiro[1]
[1]SENAI CIMATEC University Center; Salvador, Bahia Brazil

Object Detection (OD) is one of the most critical tasks in 2D image processing. The researchers proposed multiple math models and frameworks based on Deep Convolutional Networks, such as R-CNN, SSD, and YOLO are the most common. Generative Adversarial Nets (GAN) represent a prominent field of study in machine learning, and it has been applied to many tasks with exciting results. This work aims to assess the potential of GANs applied to OD tasks and the proposed frameworks as a field of study. The methodology used was a systemic review of 14 papers. The conclusion shows that although OD and GANs are popular themes, there are not many developments in the intersection of both subjects. Therefore, OD with GAN-applied tasks is an excellent field to explore in future works.
Keywords: Generative Adversarial Nets. Object Detection. Deep Learning.

Introduction

Digital Images captured by cameras are very common today as the number of such devices spiked over the last years. The captured images are no longer used only for entertainment or as an art form. Instead, they have become an essential data source that can be analyzed. In such a way that cameras are one of the most common sensors and are present in smartphones, vehicles, smart house devices, city security systems, and many others. Many aspects of an image can be analyzed, and one of the most common problems to solve is object detection in 2D images. The object detection task combines two other assignments: object classification, which identifies the type of object detected, and object localization, which identifies the object's location in the image.

The object detection problem has been gaining significant attention from the academic community and increasing momentum in publications over the last few years. Viola-Jones Detectors, HOG Detector, and Deformable Part-based Models (DPM) are some object detection frameworks. However, Deep Learning techniques, and the ones based on Convolutional Neural Networks (CNN), represented a big leap in accuracy and are the most used today [1].

Today, there are two main categories of object detectors:

1. Two-Stage Object Detectors, which separate the detection tasks of classification and localization to be held by different parts of the network
2. Single Stage Object Detectors, which do the classification and localization tasks all at once.

Two-Stage Object Detectors

The Regions with CNN (R-CNN) [2] framework consists of an initial selective search that generates region proposals by close-by pixels similarities. Then each proposed region is fed into a CNN for feature extraction. The output is used as input of bound-box regression and classification Support Vector Machine (SVM) to define if there is an object in the proposed region and what class it is. The framework presented higher Average Precision (AP) than existing methods for multiple classes [2] but with a high computational cost as it performs the redundant computation in the overlapping features of the many proposed regions. [1] Fast R-CNN [3] and SPPNet [4] apply the CNN only once in the entire image and not many times in the multiple proposed regions as per R-CNN. While

Received on 12 June 2022; revised 20 August 2022.
Address for correspondence: Caio Vinicius Bertolini. R. Dr. Barreto, 203 - apto 504V - Lauro de Freitas - BA, Brazil | CEP: 42701-310. DOI 10.34178/jbth.v5i3.228.

J Bioeng. Tech. Health 2022;5(3):202-207.

Fast R-CNN uses a Region of Interest (RoI) pooling with Fully Connected (FC) layers for classification and bounding box regression [3], SPPNet uses a Spatial Pyramid Pooling (SPP) to define the regions, which also allows for different image input sizes [4].

Faster R-CNN [5] presents inference results 34 times faster than Fast R-CNN. It happens due to the introduction of Anchor boxes and a specific Fully Convolutional Network (FCN) to generate the Region of Interest (RoI), the Region Proposed Network (RPN). The RPN uses the Anchor boxes as inputs and Intersection over Union (IoU) metrics to define the RoI bounding boxes. After that, it uses the Non-Maximum Suppression (NMS) technique to define a final bounding box. The RPN is trained together with the rest of the model, and the region outputs then pass through the RoI pooling and the FC layers for classification and bounding box regression, similarly to the Fast R-CNN [5].

Single-Stage Object Detectors

You Only Look Once (YOLO) [6] method process the entire image by a CNN-based model with multiple Anchor Boxes that simultaneously predict the class and location of the bounding boxes. Then, the bounding boxes within a pre-defined threshold are subjected to NMS to define the final bounding box for each object.

Single Shot Multibox Detector (SSD) [7] also uses multi Anchor boxes. In SSD, after the image passes through the feature extractor, it then goes by Multi-scale feature layers, which decrease in size at each step and allow predictions of detections at multiple scales at once. This ability to detect in multi-resolution and scales is the main contribution of SSD and presents an advantage in average precision compared to YOLO. The last step of SSD is NMS which eliminates the overlapping bounding boxes [1,7].

Generative Adversarial Networks (GAN)

GAN [8] is a method of a generative network using deep learning. It is divided into two parts:

A. The generator, which is an unsupervised model that takes random data input, usually from Gaussian distribution. It aims to deliver a result that can convince discriminator models that it is real.
B. The discriminator, a supervised model trained with actual data from a dataset and fake data from the generator, has the objective of distinguishing one from the other.

Both models are trained together using a loss function defined by the authors as a two-player minimax game below, where G is the generator and D is the discriminator:

$$min_G max_D V(D, G) = E_{x \sim p_{data}(x)}[\log D(x)] + E_{z \sim p_z}[\log\left(1 - D\left(G(z)\right)\right)] \quad (1)$$

The GAN model is then trained until the generator and discriminator models find an equilibrium. So the discriminator can no longer identify the difference between data generated by the generator and real data.

GAN is a fascinating field of work rapidly evolving and applies to many different problems and experiments [9].

Objective

This work aims to do a systematic review of published and pre-print works that apply GANs to image 2D object detection tasks, alone or in combination with existing object detection methods, and compare their methods and frameworks. The results of this review will evaluate the viability of GAN applied to OD tasks as a potential field of study and applied in future works.

Materials and Methods

We searched many academic and Artificial Intelligence community databases to present a comprehensive overview of GAN applications for 2D Image object detection [10-13].

Since it is a rapidly evolving field, many papers are yet to be published or as pre-prints. Therefore,

the keywords used for the search were: GAN, Object, Detection, Generative, and Adversarial. From the many works found, we first filtered the results by their titles, eliminating all the ones that were not relevant, and then by their abstracts. Since this work focuses on combining GAN techniques with object detection tasks, all papers that deviate from that were discarded, including the many papers that use GANs for data augmentation or 3D object detection.

We selected and analyzed papers with an evaluation of the methods and models used and the results obtained.

Results and Discussion

GAN and Object Detection are very popular themes in the academy. The initial search in Science Direct [11] by the keywords "generative adversarial" and "object detection" presented 851 results. We did a similar process in arXiv. og [13] with 98 initial results, Google Scholar [10] with 17,700 results, and IEEE Xplore [12] with 391 results. The results were then ranked by relevance and manually filtered according to the chosen criteria. In the end, not many works used GANs in combination with Object Detection tasks as required. Therefore, we chose 28 articles, and 14 were considered relevant to this study. We organized the results in Table 1, considering each paper's objective and proposed framework. Table 2 shows the best Mean Average Precision (map) of the models when compared to the surrogate model used in each paper and the dataset used for evaluation.

The studied authors proposed many approaches to address different tasks related to object detection, most of them to enhance the accuracy of traditional object detection frameworks. For overall object detection purposes, Wang and colleagues [14] propose an extensive trained network (teacher) for object detection and then uses a GAN for knowledge distillation from this network to a much simpler one (student) with better results in testing accuracy than the teacher model. The discriminator, in this case, checks whether the results come from the teacher or the student nets and backpropagate to the student network training. The results showed 2.8% better mAP than the surrogate studied. Prakash and Karam [15] show a baseline network trained with real data and competing with a generator net trained with augmented data. Both networks' results are then considered by a discriminator that distinguishes the results between the baseline and generator models. By the end of the adversarial training, the generator

Table 1. Selected papers proposed frameworks.

Method Objective	Paper	Proposed Framework
Overall Object Detection	[14]	SSD + GAN for Knowledge Distillation
	[15]	Competitive Object Detection Networks
Pedestrian Detection	[16]	GAN for synthetic data creation
	[17]	DCGAN + SSD
	[18]	DCGAN + SSD
	[19]	DCGAN + SSD
Small-Object Detection	[20]	Faster R-CNN + GAN
	[21]	GAN + CNN + SSD/Faster R-CNN
	[22]	CNN + ResNet + GAN
Unsupervised Bounding Box Detection	[23]	CNN + GAN + Reinforcement Learning
	[24]	Dilated CNN + GAN with Mask Mean Loss
	[25]	Encoder + Conditional GAN

Table 2. Best mAp(%) of models *vs*. Surrogate Model.

Paper	Best mAp(%) found when compared with Surrogate model and dataset	Surrogate model	Dataset used for evaluation
[14]	2.8	ResNet50	Pascal VOC 2007
[15]	2.56	SSD300	Pascal VOC 2007
[16]	Not Applicable	Not Applicable	Not Applicable
[17]	45.2	SSD	CIFAR-10/100
[18]	39.4	SSD	VOC
[19]	Not Applicable	Not Applicable	Not Applicable
[20]	19.47	Faster R-CNN	Tsinghua-Tencent 100K
[21]	25.1	FRCNN	COWC Dataset
[22]	60	Faster R-CNN (Small Objects)	Tsinghua-Tencent 100K
[23]	Not Applicable	Not Applicable	Not Applicable
[24]	5.37	[23]	Car (Stanford)
[25]	2.6	WCCN VGG16	VOC2007

model can fool the discriminator, and it is then used for inference and testing, presenting 2.56% mAp better than the baseline model. Navidan and colleagues [9] showed that synthetic data creation is a trendy application of GAN. Huang and Hamanan [16] used GAN to generate real-like pedestrian images from synthetic data generated by a game engine. It produces images of pedestrians in unusual scenarios and positions, helping traditional object detection models to improve their ability to detect pedestrians. Dinakaram and colleagues [17-19] used another known and proven ability of Deep Convolutional GAN: an image resolution improvement. All three works combine the GAN's image resolution improvement with SSD to increase its pedestrian detection capabilities in different sizes and distances. This framework can be applied in many scenarios and object classes with significant improvements in the map (Table 2). Small-Object Detection appears as an exponent application for GANs. In Huang and colleagues [20], a GAN is embedded into a Faster R-CNN Network to generate residual representations of small objects to be similar to the ones of big objects, which improves the detection ability of small objects compared to a vanilla Faster R-CNN network. Results show around 19.5% better performance in small object detection than a regular Faster R-CNN. Rabbi and colleagues [21] created a framework that uses a GAN to create high-resolution images from low-resolution images as input. The discriminator compares a real high-resolution image to the generated image. It then uses a different CNN to detect edges and improve the resolution to finally use SSD or a Faster R-CNN to detect the objects, improving the mAP performance by 25.1%. Li and colleagues [22] approach the small object detection task similarly to Huang and colleagues [20] by using region proposals. The authors presented a perceptual GAN architecture where the generator creates super-resolved representations of small objects supervised by the discriminator. The framework also uses a residual network in the generator to carry on the small object representation to be added to the generator's last part and create super-resolved features. The discriminator also has a detection and classification

branch to generate the bounding boxes and infer object class, showing 60% better mAP for small object detection than a Faster R-CNN. One of the most exciting characteristics of GANs is the ability to learn tasks in an unsupervised way. Halici and colleagues [23] and Jang and colleagues [24] use this to identify bounding boxes. Furthermore, Halici and colleagues [23] used it to differentiate images generated by the network from the same images generated in a previous network loop iteratively. So, the discriminator output is used as reinforcement learning input to the model. Jang and colleagues [24] approach the problem differently. The generator creates a black mask while the discriminator compares the generated image to the ground truth. The training stops once the generator can fool the discriminator. Diba and colleagues [25] proposed a novel ranking-discriminator network to verify the object class produced by a conditional GAN network trained with inputs from an original image representation created by an encoder network. The framework uses image-proposed regions also to identify the bounding box in a weakly supervised manner. Unsupervised OD using GANs is still a substantially unexplored field with many challenges. Although Halici and colleagues [23] and Jang and colleagues [24] propose novel approaches to this task, the results obtained are far from a state-of-the-art OD framework and do not explore Multi-Object Detection.

Conclusion

In this work, 14 papers were selected based on specific search criteria of object detection frameworks that utilize GANs in their methods. All the analyzed works proposed different approaches to the problem. Some attack object detection as a generic problem, while others have chosen to do specific tasks such as small objects or pedestrian detectors. Among the proposed frameworks, GAN was used as many tools: image enhancement, data generation, and knowledge distillation. We conclude that GAN's application to the Object Detection task does not have a preferred framework among the academic and Artificial Intelligence communities; besides, all the evaluated papers showed inspiring results. However, it presents a promising field of study to be developed in future works. In this study, we also brought up the relatively unexplored potential of unsupervised OD using GAN. It is specifically exciting if we consider the many GAN frameworks that have been proposed with multiple applications and exciting results.

References

1. Zou Z,et al. Object detection in 20 years: A survey. arXiv preprint arXi 2019;1905.05055.
2. Girshick R, et al. Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. IEEE Conference 2014:580-587.
3. Girshick R. Fast r-cnn. In: Proceedings of the IEEE international conference on computer vision. IEEE Conference 2015:1440-1448.
4. Kaiming HE, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE transactions on pattern analysis and machine intelligence IEEE 2015;37(9):1904-1916.
5. Ren S,et al. Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in Neural Information Processing Systems. 2015;28:91-99.
6. Redom J, et al. You only look once: Unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. IEEE 2016:779-788.
7. Liu W,et al. Ssd: Single shot multibox detector. In: European conference on computer vision. Springer, Cham. 2016:21-37.
8. Goodfellow I, et al. Generative adversarial nets. ACM Communications 2014:27.
9. Navidan H, et al. Generative Adversarial Networks (GANs) in networking: A comprehensive survey & evaluation. Computer Networks 2021:108-149.
10. Google Scholar. Initial Page. Available at: <https://scholar.google.com/>. Access in: Jul 25th, 2021.
11. Science Direct. Initial Page. Available at: < https://www.sciencedirect.com/>. Access in: Jul 25th, 2021.
12. IEEE Xplore. Initial Page. Available at: <https://ieeexplore.ieee.org/Xplore/home.jsp >. Access in: Jul 25th, 2021.
13 arXiv.og. Initial Page. Available at: < https://arxiv.org/>. Access in: Jul 25th, 2021.
14. Wang W, et al. Gan-knowledge distillation for one-stage object detection. IEEE Access 2020;8:60719-60727.

15. Prakash CD, Karam LJ. It GAN DO better: GAN-based detection of objects on images with varying quality. arXiv preprint arXiv 2019;1912.01707.
16. Huang S, Ramanan D. Expecting the unexpected: Training detectors for unusual pedestrians with adversarial imposters. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2017:2243-2252.
17. Dinakaran RK, et al. Deep learning based pedestrian detection at distance in smart cities. In: Proceedings of SAI Intelligent Systems Conference. Springer, Cham, 2019:588-593.
18. Dinakaran R, et al. Distant pedestrian detection in the wild using single shot detector with deep convolutional generative adversarial networks. In: 2019 International Joint Conference on Neural Networks (IJCNN) IEEE 2019:1-7.
19. Dinakaran R, Zhang L, Jiang R. In-vehicle object detection in the wild for driverless vehicles. In: Developments of Artificial Intelligence Technologies in Computation and Robotics: Proceedings of the 14th International FLINS Conference (FLINS 2020) 2020:1139-1147.
20. Huang W, Huang M, Zhang Y. Detection of traffic signs based on combination of GAN and faster-RCNN. In: Journal of Physics: Conference Series. IOP Publishing 2018:012159.
21. Rabbi J, et al. Small-object detection in remote sensing images with end-to-end edge-enhanced GAN and object detector network. Remote Sensing 2020;12(9):1432.
22. Li J, et al. Perceptual generative adversarial networks for small object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. IEEE 2017:1222-1230.
23. Halici E, Alatan AA. Object localization without bounding box information using generative adversarial reinforcement learning. In: 2018 25th IEEE International Conference on Image Processing (ICIP). IEEE 2018:3728-3732.
24. Jang H, et al. Generative object detection: Erasing the boundary via adversarial learning with mask. In: 2019 IEEE 2nd International Conference on Information Communication and Signal Processing (ICICSP). IEEE 2019:495-499.
25. Diba A, et al. Weakly supervised object discovery by generative adversarial & ranking networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops 2019.