Detecting objects in drone imagery: a brief overview of recent progress

Saša Sambolek[*], Marina Ivašić-Kos [**]
[*] High school Tina Ujevića, Kutina, Croatia
[**] University in Rijeka, Department of Informatics, Rijeka, Croatia
sasa.sambolek@gmail.com, marinai@uniri.hr

Abstract - **Detecting objects on unmanned aerial vehicle (Drones) imagery is a challenging task and an under-researched problem that lately receives more and more attention in the research community. When shooting with a drone, not only the weather and the light conditions change, but the shooting height and angle change as well because the position of the camera is not fixed during shooting. The paper aims to describe the possibility of using drones in search and rescue operations and to provide a comprehensive overview of the area related to the detection of persons in drone imagery. The paper includes a description of publicly available datasets and a comparison of the state-of-the-art models for person detection in drone recordings and ends with a proposal for future research.**

Keywords – drone dataset; object detection; CNN

I. Introduction

With the development of technology, unmanned aerial vehicles (UAVs, drones) equipped with cameras find their application in industry, agriculture, surveillance, and search and rescue operations. Detecting objects on drone images is an extremely useful and still under-researched problem.

When flying at low altitudes, drone records more details on objects of interest, while larger altitudes cover a larger area. Detecting objects in drone imagery creates greater challenges than traditional panorama detection. One of the reasons is the change in shooting height which significantly affects the size of the desired object or change of the shooting angle [1]. In a single video, in a very short time, a drone can record an object from the front, side, or a bird's eye view. Changes in lighting (day, night) and weather (sunny, cloudy, foggy, or rainy) drastically affect the visibility and display of an object [1]. In addition to all of the above, the challenge in detecting and monitoring is also posed by the rapid movements of the camera, the occlusion, and the relative movement between the camera and the object.

As objects captured by a drone are often too small to be detected by the human eye on a drone control screen, object detection needs to be automated. In recent years, considerable progress has been made in detecting objects using deep learning (convolutional neural networks). The most popular deep learning-based detectors are Faster R-CNN [1], SSD [4], YOLO [5], and RetinaNet [6] trained on datasets like PASCAL VOC [7] or MS COCO [8]. However, it turns out that they are not equally successful when applied to drone-recorded images.

In order to improve the results of object detection on drone imagery, it is necessary to include these images in a training set. However, until recently, there were no publicly available datasets recorded by the drones. With datasets like Campus [9], UAV123 [10], CARPK [11], Okutama-action [12], UAVDT [13] and VisDrone [14], images taken with the drones are available, but there are tailored to specific issues such as parking control, traffic monitoring, or movement of people across pedestrian zones.

This paper aims to provide a simple but comprehensive overview of object detection on imagery recorded by drones to study existing databases and models that could be used in the detection of persons in search and rescue operations.

During the search and rescue operation, it is important to find the missing person as quickly as possible, as the survival of the missing person declines exponentially over time [15]. The weather and light conditions vary greatly between different search and rescue operations, which is an additional challenge in detecting a missing or injured person.

Today, almost all search and rescue services have integrated the use of drones in their search.

This is due to the increasing availability of drones with quality high-resolution cameras. It takes approximately 25 seconds for a video analyst to detect a victim on a drone recording [20]. The benefit of video analytics is knowing the context of the image and predicting where the person may be based on previous experiences, but the analyst focuses only on a small portion of the image so the assistance of an automated detector can be of great use.

The rest of the paper is organized as follows: in Section II. we will present public available drones datasets. An overview of methods using drones in search and rescue missions is given in Section III with an overview of the state-of-the-art object detector algorithms in drone imagery. The paper ends with a conclusion and a proposal for future research.

II. Public Available Drones Datasets

A comparison of publicly available sets of images taken with a drone and prepared for deep learning tasks is given in Table 1.

TABLE 1. COMPARISON OF PUBLICLY AVAILABLE DRONE DATASETS

	Campus	UAV123	CARPK	Okutama-action	UAVDT	VisDrone
Year	2016	2016	2017	2017	2018	2018
D			x	x	x	x
S		x			x	x
M	x				x	x
Frames	929,5k	112 578	1 448	77 365	80 000	189 473
Boxes	19,5k	110k	90k	422,1	841,5k	2 500k
Categories	6	6	1	1	1	10
Resolution	1400x1904	1280x720	1280x720	3840x2160	1080x540	3840x2160

D – detection, S – single object tracking, M – multiple object tracking, 1k = 1 000

A. *Campus*

The Campus is a large dataset containing images and videos of different classes such as pedestrians, bicyclists, cars, skateboarders, golf carts, buses, taken inside the campus from a bird's eye view (see

Figure *1*. a). The footage was taken with a 4K drone-mounted camera (3DR solo) flying at a height of approximately 80 meters. The dataset contains about 19,000 objects at a resolution of 1400 x 1904 px, [9].

B. *UAV123*

The UAV123 dataset contains a set of scenes ranging from urban landscapes, roads, fields, and beaches with objects such as cars, trucks, boats, and persons. Persons are additionally tagged for object tracking. Activities such as walking, cycling, swimming, car driving are also labeled.

The data are divided into 3 sub-groups [10]:

- 103 video clips taken with the DJI S1000 drone tracking different objects between 5m and 25m in height, 720p, and 4K resolution at 30 and 96 fps.
- 12 videos shot with images of lower quality and resolution
- 8 synthetic video clips recorded using a drone simulator of the Unreal4 Game Engine.

C. *CARPK*

The CARPK is according to authors [11], the first and largest database of drone recordings that supports object counting, and provides the bounding box annotations. More specifically, there are 89 777 tagged cars on the dataset. The cars were shot by drones in four different parking lots. The dataset is tailored to deep learning algorithms for object count and localization scenarios.

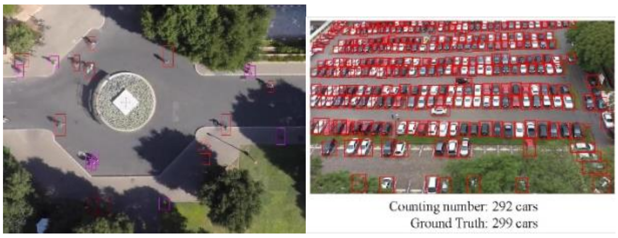

Figure 1. Examples of the scenes captured in a) Campus, b) CarPK

D. *Okutama-action*

The Okutama-action database contains 43 fully labeled drone video clips for training and testing models when detecting multiple simultaneous actions within different categories (reading, handling, carrying different items), [11].

The videos were recorded using two DJI Phantom 4 drones in 4K baseball court and at 30 fps at 10m to 45m height, while the camera angle is 45 or 90 degrees. The dataset for each video contains metadata such as camera angle, speed, and height. The shots were taken with two different lighting conditions (sunny and cloudy)

E. UAVDT

The UAVDT consists of 100 drone videos in multiple urban locations such as streets, squares, intersections, etc [11]. The videos were shot at 1080 x 540 px with 30 fps in different weather conditions (day, night, fog), and in three different altitude ranges (low: 10m to 30m, medium: 30m to 70m high: more than 70m) and different camera views (front view, side view, and bird's eye view). An example of the scene captured in the UAVDT dataset is given in

Figure 2.a.

F. VisDrone

VisDrone is a set of data shot in different scenes focusing on four basic problems in the field of computer vision (object detection in images, object detection in videos, single object tracking and multiple object tracking).

The dataset consists of 263 video clips and an additional 10.209 images [11]. Videos/images were recorded on different drone platforms (DJI Mavic, DJI Phantom Series 3, 3A, 3SE, 3P, 4, 4A, 4P) in 14 different cities in China. The dataset covers different weather and light conditions of maximum video resolution (3840 x 2160 px) and images (2000 x 1500 px).

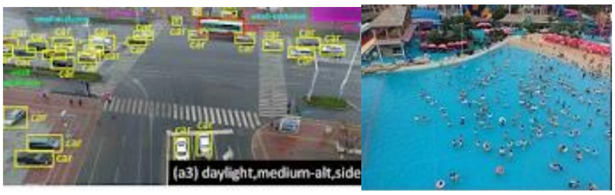

Figure 2. Examples of the scenes captured in a) UAVDT, b) VisDrone

Each of the datasets presented here is important for the development of computer vision research in the field of UAV images. However, it is clear from the descriptions of each image database and examples that they are intended for a specific task and tailored to a particular problem. For a specific problem, such as searching and rescuing people, there are missing appropriate scenarios where people in non-standard poses appear (e.g. injured persons during a fall), so they need to be recorded and included in the set.

III. Computer Vision Tasks in Search and Rescue Operations

Detection of people in images and videos plays a significant role in various applications, but in this section, we focus on search and rescue applications using drone recordings.

The search and rescue problem can be divided into four application areas: in combat, on water, in urban and non-urban areas [16]. The use of drones in search and rescue operations has been discussed in [18][19][20] [17]. In this review, we will focus on non-urban areas and water areas.

In [20] image segmentation and contrast enhancement were applied and then convolution neural networks (multiple single shot detector SSD) for the detection of persons ranging from 5 to 50 px, on drone imagery. They also used a 3D game editor to generate synthetic search and rescue datasets.

The Inception model with the Support Vector Machine (SVM) classifier is used in [21] for detecting people trapped in an avalanche using drone imagery. In [23] the focus is on detecting humans on sea recorded using an unmanned aerial vehicle equipped with a multi-spectral camera. A modified MobileNet convolution neural network architecture is used for detection.

In [24], authors have developed a system for detecting people and action recognition on the Okutama-action dataset while calculating GPS locations. For object detection, a model that was upgraded to MobileNetv2 and called POINet was used. Another example of the use of GPS signals in search and rescue actions is given in [25]. The assumption is that the injured person has a mobile device switched on, so a GSM radio signal of the mobile device is used to log the position of the injured person from the strength of the single and GPS position of the drone.

A platform for the detection of persons in water with the Tiny YOLO V3 architecture integrated on the NVIDIA Jetson X1 computer was introduced at [26]. The model was trained on the COCO dataset and swimmer's custom dataset recorded with a drone equipped with a GoPro camera in HD resolution. For the detection of sea surface objects, the use of a drone thermal camera and a real-time onboard algorithm was proposed in [27] to detect and track objects on the ocean surface.

The strategy of using semi-supervised and supervised machine learning approaches for the classification of aerial imagery and object detection along with the suggestion of hardware and software architecture for the UAV platform is given in [28]. An algorithm for planning a search path for a UAV and using unmanned ground vehicles (UVG) to verify the identity of an object detected by the UAV is given in [29]. In [30] the authors classify drone imagery on human and non-human images and provide classification results using several CNN architectures. According to [31], it was the first paper that applies multiple object visual tracking to aerial imagery for search and rescue purposes, invariant to scale, translation and rotation, and with the ability to re-identify persons. Person detection is based on color and depth data and the use of the Human Shape Validation Filter that uses the locations of human joints obtained from the Convolutional Pose Machine [32]. The purpose of the filter is to study the shape of the human skeleton on detections to avoid false detections.

According to the results of the Vision Meets Drones competition, VisDrone 2019 [33], the Cascade R-CNN [34] model and models derived from it are most commonly used to detect objects such as pedestrians, cars and bicycles in large-scale benchmark dataset covering a wide range of aspects including location (taken from 14 different cities), environment (urban and country), objects (pedestrian, vehicles, bicycles, etc.), and density (sparse and crowded scenes).

Cascade R-CNN is a multi-stage object detector framework, which aims to increase the quality of detection by constantly increasing the intersection over union (IoU) thresholds [35]. Cascade R-CNN was used in different applications including agricultural, aerial photography, fast delivery, and surveillance, followed by CenterNet [36] and RetinaNet [37].

CenterNet is a one-stage highly efficient detector for exploring the visual patterns within each bounding box. For detecting an object, this approach uses a triplet, rather than a pair, of keypoints. Paying attention to the center information, RetinaNet has a feature pyramid network (FPN) [37] attached to its backbone to generate multi-scale pyramid features. Then, pyramid features go into classification and regression branches, whose weights can be shared across different levels of the FPN. The focal loss is applied to compensate for the accuracy drop, which improves performance. The most used detectors in the VisDrone competition, as the backbone mainly use ResNet-101, ResNet 101, ResNet 50, and SEResNeXt50.

The performance that object detectors achieve on images captured with a drone is much lower than that achieved on images that are not a bird's eye view in different application domains [38, 40, 40]. The top three detectors in the VisDrone 2019 competition (DPNet-ensemble, RRNet [41], and ACM-OD) in the image detection category reach an average precision (AP) of about 29% with an IoU> 50%. For person detection a maximum of 16% AP is achieved (BetterFPN, 16.45% AP, DPNet ensemble [35], 15.97% AP, ACM-OD [35], 15.50% AP).

Slightly lower object detection results than in the case of images were achieved in the object detection category on video [44]. The three best results were achieved by the following algorithms: DBAI-Det with 29.22% AP, AF-SRNet with 24.27% AP, and HRDet+ with 23.03% AP. As in the case with the images, positive detection was counted if IoU is greater than or equal to 50%. In the case of detection of persons, the best results were achieved with DBAI-Det, VCL-CRCNN, and AFSRNet, with pedestrian detection results being different from the detection of a person in general.

The success of the best algorithms is attributed to the combination of many recently proposed powerful networks, including DCNv2 [45], FPN, and Cascade R-CNN, and detection performance is significantly enhanced by the benefits of anchor-based RetinaNet and anchorless FSAF.

IV. CONCLUSION

The paper provides an overview of the object detection on imagery recorded by unmanned aerial vehicles (drones). The first part of the paper gives an overview of the current state of publicly available datasets with their characteristics and appropriate tasks. The following section shows the research activities using drones in search and rescue operations and computer vision methods for missing person detection. Finally, the models that currently show the best detection results on images made by unmanned systems are listed.

In future work, it is necessary to create a dataset of drone recordings for better detection of injured persons. Such a dataset would contain people in atypical poses that are not contained in existing datasets. Combining knowledge transfer from existing datasets and the new custom set, it is necessary to test the state of the art models and analyze their performance in a new set of scenes characteristic for search and rescue operations. If necessary, adaptation and enhancement of existing models will be proposed to achieve the best possible detection results for disabled and missing persons in non-urban and off-water areas.

ACKNOWLEDGMENTS

This research was supported by Croatian Science Foundation under the project IP-2016-06-8345 “Automatic recognition of actions and activities in multimedia content from the sports domain” (RAASS) and the project IP-2018-01-7619 “A Knowledge-based Approach to Crowd Analysis in Video Surveillance (KACAVIS) and by the University of Rijeka (project number 18-222).

REFERENCES

[1] S. Sambolek, M. Ivasic-Kos, Detection of Toy Soldiers Taken from a Bird’s Perspective Using Convolutional Neural Networks, ICT Innovations 2019, Ohrid. Springer Communications in Computer and Information Science

[2] M. Kristo, M. Ivašić-Kos, Thermal Imaging Dataset for Person Detection; Proceedings of 42nd International ICT Convention – MIPRO 2019, Opatija, Hrvatska: Mipro, 2019. str. 1316-1321

[3] S. Ren, K. He, R. Girshick, J. Sun, “Faster R-CNN: Towards real-time object detection with region proposal networks,” in Advances in neural information processing systems, 2015, pp. 91-99.

[4] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, A. C. Berg, “SSD: Single shot multi-box detector,” in European conference on computer vision, Springer, Cham, 2016, pp. 21-37.

[5] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, “You only look once: Unified, real-time object detection,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 779-788.

[6] T. Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, “Focal loss for dense object detection,” in Proceedings of the IEEE international conference on computer vision, 2017, pp. 2980-2988.

[7] M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, A. Zisserman, “The pascal visual object classes challenge: A retrospective,” International journal of computer vision, 2015, 111(1), 98-136.

[8] T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: Common objects in context. In ECCV, 2014.

[9] A. Robicquet, A. Sadeghian, A. Alahi, S. Savarese, “Learning social etiquette: Human trajectory understanding in crowded scenes,” in European conference on computer vision, Springer, Cham, 2016, pp. 549-565.

[10] M. Mueller, N. Smith, B. Ghanem, “A benchmark and simulator for UAV tracking,” in European conference on computer vision, Springer, Cham, 2016, pp. 445-461.

[11] M. R. Hsieh, Y. L. Lin, W. H. Hsu, “Drone-based object counting by spatially regularized regional proposal network,” in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 4145-4153.

[12] M. Barekatain, M. Martí, H. F. Shih, S. Murray, K. Nakayama, Y. Matsuo, H. Prendinger, “Okutama-action: An aerial view video dataset for concurrent human action detection,” In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2017, pp. 28-35.

[13] D. Du, Y. Qi, H. Yu, Y. Yang, K. Duan, G. Li, Q. Tian, “The unmanned aerial vehicle benchmark: Object detection and tracking,” in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 370-386.

[14] P. Zhu, L. Wen, X. Bian, H. Ling, Q. Hu, “Vision meets drones: A challenge,” arXiv preprint arXiv:1804.07437, 2018.

[15] R. J. Koester, Lost Person Behavior: A Search and Rescue. DBS Productions LLC, 2008.

[16] S. N. A. M. Ghazali, H. A. Anuar, S. N. A. S. Zakaria, Z. Yusoff, “Determining position of target subjects in maritime search and rescue (MSAR) operations using rotary-wing unmanned aerial vehicles (UAVs),” in 2016 International Conference on Information and Communication Technology (ICICTM), IEEE, 2016, pp. 1-4.

[17] P. Doherty, P. Rudol, “A UAV search and rescue scenario with human body detection and geolocalization,” in Australasian Joint Conference on Artificial Intelligence, Springer, Berlin, Heidelberg, 2007, pp. 1-13.

[18] M. A. Goodrich, B. S. Morse, C. Engh, J. L. Cooper, J. A. Adams, “Towards using unmanned aerial vehicles (UAVs) in wilderness search and rescue: Lessons from field trials,” Interaction Studies, 2009, 10(3), 453-478.

[19] S. Waharte, N. Trigoni, “Supporting search and rescue operations with UAVs,” in 2010 International Conference on Emerging Security Technologies, IEEE, 2010, pp. 142-147.

[20] C. A. Baker, S. Ramchurn, W. T. Teacy, N. R. Jennings, “Planning search and rescue missions for UAV teams,” in Proceedings of the Twenty-second European Conference on Artificial Intelligence, IOS Press, 2016, pp. 1777-1778.

[21] K. Yun, L. Nguyen, T. Nguyen, D. Kim, S. Eldin, A. Huyen, E. Chow, “Small target detection for search and rescue operations using distributed deep learning and synthetic data generation,” in Pattern Recognition and Tracking XXX (Vol. 10995, p. 1099507), International Society for Optics and Photonics, 2019.

[22] M. Bejiga, A. Zeggada, A. Nouffidj, F. Melgani, “A convolutional neural network approach for assisting avalanche search and rescue operations with UAV imagery,” Remote Sensing, 2017, 9(2), 100.

[23] A. J. Gallego, A. Pertusa, P. Gil, R. B. Fisher, “Detection of bodies in maritime rescue operations using unmanned aerial vehicles with multispectral cameras,” Journal of Field Robotics, 2019, 36(4), 782-796.

[24] R. Geraldes, A. Gonçalves, T. Lai, M. Villerabel, W. Deng, A. Salta, H. Prendinger, “UAV-based situational awareness system using deep learning,” IEEE Access, 2019, 7, 122583-122594.

[25] S. O. Murphy, C. Sreenan, K. N. Brown, “Autonomous unmanned aerial vehicle for search and rescue using software-defined radio,” in 2019 IEEE 89th Vehicular Technology Conference VTC2019-Spring, 2019, pp. 1-6. IEEE.

[26] E. Lygouras, N. Santavas, A. Taitzoglou, K. Tarchanidis, A. Mitropoulos, A. Gasteratos, “Unsupervised human detection with an embedded vision system on a fully autonomous UAV for search and rescue operations,” Sensors, 2019, 19(16), 3542.

[27] F. S. Leira, T. A. Johansen, T. I. Fossen, “Automatic detection, classification and tracking of objects in the ocean surface from UAVs using a thermal camera,” in 2015 IEEE aerospace conference, IEEE, 2015, pp. 1-10.

[28] J. Sun, B. Li, Y. Jiang, C. Y. Wen, “A camera-based target detection and positioning UAV system for search and rescue (SAR) purposes,” Sensors, 2016, 16(11), 1778.

[29] Z. Kashino, G. Nejat, B. Benhabib, “Aerial wilderness search and rescue with ground support,” Journal of Intelligent & Robotic Systems, 2019, 1-17.

[30] T. Marasović, V. Papić, “Person classification from aerial imagery using local convolutional neural network features,” International Journal of Remote Sensing, 2019, 1-19.

[31] A. Al-Kaff, M. J. Gómez-Silva, F. M. Moreno, A. de la Escalera, J. M. Armingol, "An appearance-based tracking algorithm for aerial search and rescue purposes," Sensors, 2019, 19(3), 652.

[32] S. E. Wei, V. Ramakrishna, T. Kanade, Y. Sheikh, "Convolutional pose machines," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4724-4732.

[33] D. R. Pailla, "VisDrone-DET2019: the vision meets drone object detection in image challenge results, 2019.

[34] Z. Cai, N. Vasconcelos, "Cascade R-CNN: Delving into high-quality object detection," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 6154-6162.

[35] M. Burić, M. Pobar, M. Ivasic-Kos, Adapting YOLO Network for Ball and Player Detection; Proceedings of the 8th International Conference on Pattern Recognition Applications and Methods - Volume 1: ICPRAM, Prag, Češka: SciTePress, 2019. str. 845-851

[36] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, Q. Tian, "Centernet: Keypoint triplets for object detection," in Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 6569-6578.

[37] T. Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, "Feature pyramid networks for object detection," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2117-2125.

[38] M. Ivasic-Kos, M. Krišto, M. Pobar, Person Detection in Thermal Videos Using YOLO; Proceedings of SAI Intelligent Systems Conference IntelliSys 2019: Intelligent Systems and Applications, Cham: Springer, 2019. str. 254-267

[39] M. Pobar, M. Ivasic-Kos, Detection of the leading player in handball scenes using Mask R-CNN and STIPS, Proc. SPIE 11041, Eleventh International Conference on Machine Vision (ICMV 2018), Muenchen: SPIE, 2018

[40] M. Ivasic-Kos, M. Pobar; Building a labeled dataset for recognition of handball actions using Mask R-CNN and STIPS, 7th European Workshop on Visual Information Processing EUVIP, Tampere, Finska: IEEE, 2018. str. 1-6

[41] C. Chen, Y. Zhang, Q. Lv, S. Wei, X. Wang, X. Sun, J. Dong, "RRNet: A hybrid detector for object detection in drone-captured images," in Proceedings of the IEEE International Conference on Computer Vision Workshops, 2019.

[42] K. Chen, J. Wang, J. Pang, Y. Cao, Y. Xiong, X. Li, Z. Zhang, "MMDetection: Open MMLab detection toolbox and benchmark, 2019, arXiv preprint arXiv:1906.07155.

[43] B. Singh, M. Najibi, L. S. Davis, "SNIPER: Efficient multi-scale training," in Advances in Neural Information Processing Systems, 2018, pp. 9310-9320.

[44] P. Zhu, D. Du, L. Wen, X. Bian, H. Ling, Q. Hu, T. Peng…, "VisDrone-VID2019: "The vision meets drone object detection in video challenge results," 2019.

[45] X. Zhu, H. Hu, S. Lin, J. Dai, "Deformable convnets v2: More deformable, better results," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 9308-9316.