A UAV Object Detection Benchmark for Vision-assisted Powerline Element Inspection

Emmanouil Patsiouras
Department of Informatics
Aristotle University of Thessaloniki
Thessaloniki, Greece
emmanoup@csd.auth.gr

Vasileios Mygdalis
Department of Informatics
Aristotle University of Thessaloniki
Thessaloniki, Greece
mygdalisv@csd.auth.gr

Ioannis Pitas
Department of Informatics
Aristotle University of Thessaloniki
Thessaloniki, Greece
pitas@csd.auth.gr

***Abstract*—Powerline inspection operations involve capturing and inspecting visual footage of powerline elements in electric transmission infrastructures. Current technological advantages in the areas of robotics and machine learning are towards enabling the utilization of completely autonomous Unmanned Aerial Vehicles (UAVs) to carry out such tasks. One of the tasks to be addressed is the robust, precise, and fast powerline object detection problem. To this end, UAVs are required to perform visual object detection autonomously, with high accuracy and fast algorithm execution speed, for providing image regions of interest to be inspected by humans or even be used as input for autonomously controlling the UAV/camera. However, the limited computational resources of the on-board devices of such systems heavily affect the type of neural network architectures that can potentially be deployed. In this work, we study state-of-the-art object detectors in an attempt to find an acceptable trade-off between detection accuracy and inference speed that will allow the exploitation of UAVs for autonomous powerline inspection purposes. To this end, we publicly release a powerline inspection dataset and state a benchmark evaluation with recently proposed object detectors based on deep learning.**

Keywords—deep learning, object detection, powerline inspection, unmanned aerial vehicles

I. INTRODUCTION

Powerline inspection operations involve gathering detailed video footage from elevated positions close and around powerline elements, such as towers, insulators and dumpers. This operation is usually carried out by helicopters thus, is accompanied by high logistic costs and complicated planning requirements (mostly linked with the determination of safe take-off/landing spots) and personnel involved (pilots, cameramen). Captured video footage is thereby inspected by appropriately trained human workers. Over recent years, camera-equipped Unmanned Aerial Vehicles (UAVs) have been successfully utilized in many visual tasks such as sports cinematography [1], media production [2], road traffic surveillance [3], and search and rescue missions [4]. One such related and emerging task is the linear infrastructure inspection [5], where UAVs may potentially provide added operation value when compared to traditional inspection methods by minimizing the personnel involved and increasing the operational efficiency.

In this paper we focus on the object detection task, where the goal is to visually localize, identify and monitor specific elements and components in high-voltage transmission lines. We argue that this task poses challenges that may not be present in standard benchmark datasets such as MS-COCO [6]. More specifically, the successful detection of powerline elements requires high precision in both small and big powerline elements appearing on the same image frame (e.g., towers and insulators), robustness in difficult illumination conditions (sometimes even against the sun), and the ability to discriminate visually similar and small elements on the powerline (e.g., insulators and dumpers) while these elements may have minimal contrast and chromatic differences. In order to address the aforementioned task and the associated challenges our attention is focused on state-of-the art methods for object detection commonly employed in UAVs, including the Single-Shot Detector (SSD) [7] and the You Only Look Once (YOLO) [8] detector. To the best of our knowledge, none of the above mentioned Deep Learning-based (DL) detection methods have been deployed in online UAV powerline inspection tasks.

As already mentioned, the goal of this paper is to utilize UAVs and DL in order to detect, in an intelligent manner, elements and components most commonly found in powerline grids. Recent technological progress has led to the production of such affordable UAVs but the limited computational resources of such on-board devices are significantly narrowing the performance of any DL-based detection model hence, rendering their straightforward deployment rather challenging [9]. An additional bottleneck in deploying DL-based models for powerline inspection is the lack of publicly available datasets for the desired use case. To this end, we propose the AERIAL-CORE powerline inspection dataset for training and evaluating our studied detection models, while attempting to find an acceptable trade-off that will facilitate their immediate deployment on UAVs.

The remaining of this paper is organized as follows. Section II overviews the state-of-the-art in object detection and focuses on methods benchmarked in the proposed dataset. In Section III, we analytically describe the motivation and details of our proposed powerline inspection dataset, along with the arising difficulties in the powerline element detection application scenario. In Section IV, we present the experiments conducted and discuss the obtained results. Finally, Section V draws our conclusions.

II. OBJECT DETECTION

The task of object detection consists of classifying and localizing semantic objects of interest depicted in an image frame, in the form of rectangular Regions of Interest (ROIs),

commonly referred to as bounding boxes. In the linear infrastructure inspection application scenario, detection involves the precise localization of powerline elements and components on the power transmission lines (i.e., insulators, dumpers and towers). The most typical way of addressing these objects is through predicting a set of bounding boxes associated with class labels. Given a set of classes $\mathcal{C} = \{\mathcal{C}_i = 1, \dots, m\}$ and an image $\mathbf{x} \in \mathbb{R}^n$ the detection model $\hat{\mathbf{y}} = f(\mathbf{x}; \theta)$ predicts (assuming only one object instance) an output $\hat{\mathbf{y}} = [\hat{\mathbf{y}}_1^T, \hat{\mathbf{y}}_2^T]$ consisting of: (i) a class vector $\hat{\mathbf{y}}_1^T \in [0,1]^m$, and (ii) a bounding box parameter vector $\hat{\mathbf{y}}_2^T = [x, y, w, h]^T$ corresponding to object ROI.

State-of-the-art performing object detection methods rely on the power of deep Convolutional Neural Networks (CNNs), which are very effective in visual data classification tasks. CNN object detection essentially relies on the two-class classification of local image regions in the UAV camera video frames, of whether some specific target object is depicted or not. This problem is virtually impossible to be solved exhaustively, since almost infinite combinations of localized image areas can be derived. Therefore, only a subset of such regions is examined during a detection step. Depending on whether or not they rely on region proposals [10], CNN object detection methods fall into one of two categories, namely two-stage and single-stage detectors, respectively. Finally, a Non-Maximum Suppression (NMS) step is ultimately performed to eliminate multiple detections of the same object in overlapping regions.

Preliminary methods such as Fast R-CNN [11], relied on image processing algorithms such as Selective Search [11] for producing region proposals, which was performed prior to the extraction of CNN features on each image proposal, or after extracting features once for each image, respectively. Later on, Faster R-CNN [10] replaced the Selective Search process with Region Proposal Network (RPN), that learns to predict object proposals using an optimization procedure, based on the outcome of regions proposals that were finally selected or discarded by its predecessors. Despite the performance improvements, detectors based on region proposals are still not fast enough to be implemented on drones.

Perhaps the most prominent state-of-the-art object detection approaches that could achieve an adequate performance in the aforementioned application scenario, best reconciling between detection accuracy and inference speed, belong to the family of single-stage object detectors. Such single-stage object detectors, e.g., SSD [7], CornerNet [12] and YOLO [8] treat the detection task as a simple regression and classification problem by receiving an input image and simultaneously predicting class probabilities and bounding box coordinates. We focus our study on two state-of-the-art single-stage detectors, namely the SSD [7] and YOLOv4 [13] detectors, and the recently introduced Detection with Transformers (DETR) [14] method that simplifies the detection pipeline. SSD utilizes a set of predefined boxes, called anchors, of different aspect ratios and scales in order to predict the presence of an object in an image. SSD captures all the necessary computations in a single network, meaning that a single feed-forward pass of an image suffices for the extraction of multiple ROIs with coordinate and class information. In [15], SSD was pitted against a number of two-stage detection methods and amongst the findings was that when combined with Mobilenets [16] and Inception v2 [17] feature extractors, it prevailed in terms of speed at the cost, however, of lower detection precision. YOLO [8] is another family of fast single-stage object detectors. Similar to SSD, YOLO utilizes anchors as a set of fixed candidate regions in order to directly predict detections. YOLO works by dividing an input image into a standard $S \times S$ grid and for each grid cell predicts a number of bounding boxes accompanied with confidence scores and class probabilities. YOLO relies on custom backbone architectures for feature extraction. Since its initial proposal, YOLO has undergone a number of improvements with its latest version, YOLOv4 [13], achieving superior performance in both time and accuracy.

As mentioned, the above methods treat the detection task as a combination of classification and bounding box regression heavily relying on some trivial hand-crafted components such as anchor generation, and NMS for collapsing overlapping bounding boxes. However, the recently introduced DETR, views object detection as a direct set prediction problem that completely eliminates the need of any geometric priors. DETR incorporates an attention mechanism in the form of a Transformer [17] architecture in the overall pipeline. DETR initially passes an image through a backbone in order to extract feature representations which are subsequently fed into an encoder that outputs higher-level features with attention information. The decoder then takes a fixed number of learnable positional embedding as object queries and additionally attends to the output of the encoder. Finally, DETR passes the normalized output of the detector to a simple multi-layer perceptron network that predicts either a detection, meaning class plus bounding box, or a "no object" class.

III. Powerline Element Inspection

Camera-equipped UAVs deployed with efficient DL-vision based models are able to collect data and perform visual analysis for both offline and online inspection, in conjunction with a human aerial work crew, to quickly identify and repair any damaged or faulty components in the powerline. Traditionally, these inspection tasks have been carried out by human operators and helicopter-assisted surveys, sometimes days or weeks after the initial flight. Hence, UAV-enabled powerline inspection has the potential of reducing operational risks, time and costs associated with the aforementioned conventional methods. In order to provide this added value, the detection task performed on-drone needs to be carried out as efficiently and accurately as possible.

As already mentioned in the introduction section, DL-vision based powerline inspection comes with quite a few challenges. Maybe the most prominent challenge is the requirement to simultaneously detect small and big electrical components that appear from short and long shooting distances, diverse viewing angles, and with various illumination conditions. Further challenges are imposed when these objects are clutched against an indistinguishable background or when distortions are produced due to difficult lighting conditions. It should also be noted that some electrical components (e.g., dampers and insulators) visually appear in similar chromatic color ranges and have similar shapes, sometimes rendering them indistinguishable even to humans, especially in low-resolution images. Note that low-resolution images are typically fed to DL models in embedded devices, due to computational and memory limitations. Moreover, in order to accurately detect the desired powerline elements, DL-

Figure 1. Image samples of the AERIA-CORE powerline inspection dataset and elements detected by DETR detector. The dataset contains a high variety of image shots.

vision based models usually require a large amount of training data. To the best of our knowledge, there are no publicly available datasets that are big enough in order to provide satisfactory training results. Individual researcher efforts so far, such as the CPLID dataset [19] only include few hundred images without capturing realistic scenarios (they are captured from fixed distance above the powerline).

To this end, we collected the AERIAL-CORE powerline inspection dataset, consisting of 11587 RBG images, acquired from video footage of a realistic aerial inspection performed by a helicopter, operating above and around the area surrounding the power lines. The captured data are very diverse, in the sense that they include image frames from several high-transmission lines, shot from different viewing angles, and from variable distance from the powerline elements of interest. The dataset was annotated with objects most commonly found in high-voltage transmission lines, i.e. insulators, dumpers and electric towers. The photographic data have been annotated in a per-frame basis with ROIs of the aforementioned object classes using specialized software. This dataset, which will be made publicly available, was constructed in order to train and evaluate our choice of detection methods mentioned in the previous section. Table 1 shows the performed dataset split as well as some additional properties of our dataset. Examples of the dataset along with visual results of the experimented detection models are presented in Figure 1.

TABLE I. TRAINING/EVALUATION SPLIT AND ADDITIONAL PROPERTIES OF THE AERIAL-CORE DATASET. THE ANNOTATION TYPE REFERS TO ANNOTATING THE IMAGES WITH BOUNDING AS WELL AS CLASSIFICATION LABELS.

#Training images	#Test images	Size	Annotation
8422	3165	1280 x 720	BB+label

IV. EXPERIMENTAL STUDY

In this Section, we discuss the experimental setup and report results offering a speed/accuracy trade-off analysis for the benchmarked object detectors. Inspired by [15], we tested our models on different configurations by alternating components such as the feature extractor and input image resolution. For evaluation metrics, average precision (AP) and frames per second (FPS) were used to record detection accuracy and inference speed, respectively. For all models, a COCO pre-trained model was used for capturing general features, which was then fine-tuned on the AERIAL-CORE powerline inspection dataset in order to capture task specific patterns. The comprehensive results of our experimental study are reported in Table 2. For FPS measurement, we obtained results using an on-board UAV device with limited computational resources, namely an Nvidia Jetson Xavier AGX, and also a powerful desktop Nvidia 2080 Super GPU.

For our choice of feature extractors, we experimented with both heavy architectures such as the Resnet50 [20] and more lightweight ones such as the Mobilenet v2 [16] and Inception v2 [17]. All remaining standard parameters settings regarding each individual neural network (e.g., learning rate, weight decay) were set equal to the values proposed by the respective authors. In the case of SSD, when fixed input size is used, i.e., 256×448, we obtain the best recorded FPS when using Mobilenets v2 as feature extractor, i.e., 17 FPS, while also achieving the best average precision, i.e., 50.1 average precision. Regarding YOLOv4, we experimented with a single input size and one backbone, i.e., the CSPDarknet53 [21]. Although YOLOv4 achieves only 41.6 average precision, it manages to outperform every other detection method in terms of speed achieving a real-time performance of 26 FPS. Finally for DETR, we conducted experiments with Resnet50 as a feature extractor and for different input sizes. We notice that when compared to SSD with the same backbone and for the same input size, DETR outperforms SSD in speed readings achieving 12 and 8 FPS, for 256×448 and 448×796 input sizes, receptively, while vastly outperforming it in terms of detection accuracy achieving 52.4 and 56.3 average precision.

Out of these experiments, we can conclude that up to date, only YOLOv4 method has the potential of being applied in a realistic UAV scenario, due to its ability to run in real time in

the embedded platform. Another critical factor for obtaining sufficient detection accuracy, especially on small and medium objects (e.g., 32^2< area <96^2 pixels), is to leverage higher input image resolutions, as shown by the performance of SSD and DETR methods. This is due to the fact that some powerline elements may be visually indistinguishable from long distances, hence rendering low input resolutions as the most probable cause of most detection failures.

TABLE II. COMPARISON OF STATE-OF-THE-ART OBJECT DETECTORS USING DIFFERENT BACKBONES AND IMAGE RESOLUTIONS IN THE AERIAL-CORE POWERLINE INSPECTION DATASET. AVERAGE PRECISION (AP) AND FRAMES PER SECOND (FPS) WERE USED AS EVALUATION METRICS.

Model	FPS 2080/Xavier	Input Size	AP	AP_{50}
SSD Mobilenet v2	136/22	128 x 256	42.2	75.0
SSD Mobilenet v2	126/17	256 x 448	50.1	82.1
SSD Inception v2	84/13	256 x 448	48.7	80.0
SSD Resnet 50	40/9	256 x 448	48.3	73.6
SSD Resnet 50	31/6	448 x 796	52.3	79.8
YOLOv4 Dark53	91/26	256 x 448	41.6	83.5
DETR Resnet50	45/12	256 x 448	52.4	83.1
DETR Resnet50	35/8	448 x 796	56.3	86.0

V. CONCLUSION

In this work we tackled the task of powerline inspection by utilizing UAVs and DL object detection approaches. We discussed why powerline inspection is a rather challenging task, even for conventional inspection methods, and demonstrated how UAVs can help alleviate certain limitations. To further address this task, we created a powerline inspection dataset for training and evaluating the state-of-the-art object detection methods. Finally, we investigated the behaviour of those methods by offering a speed/accuracy trade-off analysis, while focusing on their performance on devices with limited computational resources.

ACKNOWLEDGMENT

This work has received funding from the European Union's Seventh Horizon 2020 research and innovation programme under grant agreement number 871479 (AERIAL-CORE). We would like to thank EDISTRIBUCIÓN Redes Digitales, S.L. for providing the aerial video footage. This publication reflects the authors' views only. The European Commission is not responsible for any use that may be made of the information it contains.

REFERENCES

[1] I. Karakostas, V. Mygdalis, A. Tefas, and I. Pitas, "Occlusion detection and drift-avoidance framework for 2d visual object tracking," Signal Processing: Image Communication, vol. 90, pp. 116011, 2021.

[2] I. Mademlis, V. Mygdalis, N. Nikolaidis, M. Montagnuolo, F. Negro, A. Messina, and Ioannis Pitas, "High-level multiple-uav cinematography tools for covering outdoor events," IEEE Transactions on Broadcasting, vol. 65, no. 3, pp. 627–635, 2019.

[3] M. Elloumi, R. Dhaou, B. Escrig, H. Idoudi, and L. A. Saidane, "Monitoring road traffic with a uav-based system," in 2018 IEEE Wireless Communications and Networking Conference (WCNC). IEEE, 2018, pp. 1–6.

[4] M. Połka, Szymon Ptak, and Łukasz Kuziora, "The use of uav's for search and rescue operations," Procedia engineering, vol. 192, pp. 748–752, 2017.

[5] A. Al-Kaff, F. M. Moreno, L. J. S. Jose, F. Garcia, D. Martin, A. Escalera, A. Nieva, and J. L. M. Garcea, "Vbii-uav: Vision-based infrastructure inspection-uav," in World Conference on Information Systems and Technologies. Springer, 2017, pp. 221–231.

[6] T.Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Doll´ar, and C. L. Zitnick, "Microsoft coco: Common objects in context," in European Conference on Computer Vision (ECCV). Springer, 2014, pp. 740–755.

[7] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in Proceedings of the European Conference on Computer Vision (ECCV). Springer, 2016, pp. 21–37.

[8] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 779–788.

[9] P. Nousi, E. Patsiouras, A. Tefas, and I. Pitas, "Convolutional neural networks for visual information analysis with limited computing resources," in Proceedings of the IEEE International Conference on Image Processing (ICIP), 2018, pp. 321–325.

[10] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in Advances in Neural Information Processing Systems (NIPS), 2015.

[11] R. Girshick, "Fast R-CNN," in Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1440-1448.

[12] H. Law and J. Deng, "Cornernet: Detecting objects as paired keypoints," in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 734–750.

[13] C. Y. Bochkovskiy, A.and Wang and H. Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," arXiv preprint arXiv:2004.10934, 2020.

[14] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," arXiv preprint arXiv:2005.12872, 2020.

[15] J. Huang, V. Rathod, C. Sun, M. Zhu, A. Korattikara, A. Fathi, I. Fischer, Z. Wojna, Y. Song, S. Guadarrama, and K. Murphy, "Speed/accuracy trade-offs for modern convolutional object detectors," in Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 7310–7311.

[16] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4510–4520.

[17] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," arXiv preprint arXiv:1502.03167, 2015.

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in Advances in Neural Information Processing Systems (NIPS, 2017, pp. 5998–6008.

[19] X. Tao, D. Zhang, Z. Wang, X. Liu, H. Zhang, and D. Xu, "Detection of power line insulator defects using aerial images analyzed with convolutional neural networks," IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2018.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE conference on Computer Vision and Pattern recognition (CVPR), 2016, pp. 770–778.

[21] C. Y. Wang, H. Y. Mark Liao, Y. H. Wu, P. Y. Chen, J. W. Hsieh, and I. H. Yeh, "Cspnet: A new backbone that can enhance learning capability of cnn," in Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2020, pp. 390-391.