Case Studies in Construction Materials 18 (2023) e01917

Contents lists available at ScienceDirect

Case Studies in Construction Materials

journal homepage: www.elsevier.com/locate/cscm

Case study

Deep learning applied to equipment detection on flat roofs in images captured by UAV

Lara Monalisa Alves dos Santos [a,*], Vanda Alice Garcia Zanoni [a], Eduardo Bedin [b], Hemerson Pistori [b,c]

[a] University of Brasilia, Distrito Federal, Brazil
[b] Dom Bosco Catholic University, Mato Grosso do Sul, Brazil
[c] Federal University of Mato Grosso do Sul, Mato Grosso do Sul, Brazil

ARTICLE INFO

Keywords:
Flat roof maintenance
Detection of equipment
Building inspection
Unmanned Aerial Vehicle (UAV)
Computer Vision
Deep Learning
CNN-based object detection methods

ABSTRACT

Maintenance on flat roofs is a complex activity. Equipment improperly positioned on flat roofs hinders the correct drainage of water and makes maintenance services more difficult. This article presents an experiment with deep learning algorithms involving 330 images acquired in 9 buildings by Unmanned Aerial Vehicle-UAV. This dataset was created by the authors to optimize decision-making for maintenance through automated processes and is being used for the first time in this article. The dataset refers to condenser equipment positioned on flat roofs and was tested in six state-of-the-art object-detection deep learning algorithms: Region-based convolutional neural networks (Faster R-CNN), Focal Loss (Retina-Net), Adaptive Training Sample Selection (ATSS), VarifocalNet (Vfnet), Side-Aware Boundary Localization (SABL) and FoveaBox (Fovea). Nine performance metrics were applied, achieving successful results by Faster R-CNN (Recall=0.93, Fscore=0.93, MAE=0.43) followed by ATSS (Precision=0.95). In a system with many variables, the target is the identification of the best algorithm capable of solving the proposed problem. In conclusion, the types of errors analyzed in detection alert to the diversity of causes related to the inherent characteristics of flat roofs that induce network confusion.

1. Introduction

Inspection is an essential task for building maintenance services, as it is the process that helps in the diagnosis of present damage, aiming at repairs and replacements that influence the performance of the building system. Rehabilitation and maintenance activities are key factors for the sustainability and service life of the building [1]. In the roofing system in buildings, the planned building inspection procedure has been neglected, making difficult the preventive verification of damage [2]. Mapping of damage in roofing systems in advance is essential to maintain the lowest levels of intervention, as long as the problems are repaired within the expected maintenance deadlines [3].

Traditional inspection procedures done manually by specialized professionals are ostensibly time-consuming, laborious and pose threats to the health and safety of inspectors, especially considering working at heights on difficult-to-access roofs and the long time spent on inspections. In addition, they are basically sensory, intrinsically visual, and subjective, which will depend on the inspector's experience [4]. Due to these disadvantages, the Architecture, Engineering, Construction, and Operation (AECO) sector invests in new

[*] Corresponding author.
E-mail address: 200054104@aluno.unb.br (L.M.A. Santos).

https://doi.org/10.1016/j.cscm.2023.e01917
Received 9 December 2022; Received in revised form 30 January 2023; Accepted 3 February 2023
Available online 9 February 2023

technologies, in order to facilitate inspection processes, reducing time, risk of accidents, and errors in data collection and processing [5].

The use of drones for capturing GPS-oriented aerial images during building inspection and for 3D modeling [6,7] has grown rapidly in the last decade [8,9]. Establishing a comparison with traditional inspection methods, the advantages of using the drone are associated with the flexibility to carry out work at height with difficult access, ease of adaptation to different operating conditions [6,10], mobility in data acquisition due to the ability to fly at different heights, with high resolution of digital images [11].

The field survey with a drone for the detection of damage is a necessary step for the creation of a dataset [12]. In the roofing system, the acquired images record the as-built condition that can be observed visually and contribute to the inventory of equipment through object recognition. Considering a dataset, the technological resources given by deep learning applied in computer vision provide the automation of damage maps based on image patterns and the monitoring of the state of conservation. In this context, this paper presents an experiment carried out to obtain a dataset of condenser equipment positioned on flat roofs and its applicability in deep learning to verify the performance of six state-of-the-art object detection algorithms.

This study was motivated by the need to identify and record equipment installed on large roofs in buildings with many built assets as in the case selected for the proposed experiment. Mainly, there is an interest in automated processes to optimize decision-making for maintenance planning. This type of equipment arranged on the roof may be deactivated, without functionality, preventing the operation of the drains or causing water puddles, which accelerates the degradation of the waterproofing blanket and causes moisture infiltration. In addition, complex buildings, such as public buildings owned by government institutions, have a centralized administration that hires service providers and purchases materials through budgets and bidding that should be based on a dataset. Audits of maintenance and conservation plans, as well as asset controls (facilities and equipment), are objects of interest in analysis and drive advances in technological research.

Among the main contributions reported in this article, the construction of an original and unpublished dataset is highlighted. Furthermore, in a complex system with many variables, the identification of the successful convolutional neural networks capable of solving the proposed problem was a challenging task. Including, the analysis of the causes that induce the network confusion was a relevant contribution.

2. Overview of deep learning for the detection of equipment positioned on flat roofs

Convolutional Neural Networks (CNN) has been applied on a large scale for image detection and classification, in addition to being effective in discovering complex structures in datasets [13]. However, in the area of Architecture, Engineering, Construction, and Operation (AECO) automated processes using Computer Vision still require a lot of research. New digital technologies associated with image acquisition gain relevance through automated processes based on machine learning, making it a more dynamic and faster task [7,8,14].

Especially in buildings, the possibility of associating the drone with the neural network is a suitable option for quick and safe roof performance inspections [15]. Algorithms applied in neural networks can be used to process measurements, detect risk situations, or evaluate existing conditions, aiming to classify the actions of protection, maintenance, or repair of buildings [16,17]. The use of an automated inspection system requires robust detection and classification algorithms using convolutional neural networks [3]. This type of detection method improves visual recognition and object detection, potentiating discoveries in large datasets [18].

To evaluate the use of drones in image capture, Silveira, Melo, and Costa [5] collected data from 167 roofs of housing. The authors compare the difficulties faced during the building inspection of the roofing system with the drone and the traditional method, based on accessibility limitations, risk of falls, and the quality of images for the identification of damages.

The authors Bown and Miller [2] researched the quality of images acquired by the drone during inspections on a roof of 9144 m^2, in a flight plan performed safely by amateur pilots in less than two hours. The authors considered that the flights were performed efficiently for the inspection process, whose image production was of sufficient quality for decision-making and maintenance. In addition, the choice of drone, the quality of image acquisition, and the type of flight are factors that help to obtain a methodological approach that eliminates the risks involved in sensory inspection, with results in a manageable format and reduced risk of collision with obstacles, facilitating inspection and reducing the time spent [2].

The use of the drone to carry out an inspection of the roof system is also addressed by Banaszek, Banaszek, and Cellmer [11] who describe the way of acquiring 165 aerial images taken at heights of 80 m, 60 m, 40 m, and 20 m. The images allowed an assessment of the state of conservation of the roofing system and the equipment present, without the need for the inspector to be physically present on the roof. The authors discuss the relationship between flight height and image quality since there are several factors that interfere with a result capable of identifying visible damage.

The authors Yudin et al. [19] used an image segmentation algorithm via Deep Fully Convolutional Network software to detect irregularities that cause water accumulation in flat roofs, through aerial photographs, making it possible to detect and measure the size and perimeter of the cavities. Rakha et al. [6] used photogrammetry, heat mapping through thermal imaging, and the use of computer vision workflows to analyze and segment thermal images, autonomously detecting damage.

Buildings that have real estate insurance need evaluators to survey the state of conservation of the roofing system, especially when it comes to accident events (earthquakes, fires, hail, windstorms). The authors Hezaveh, Kanan, and Salvaggio [3] propose to automate the evaluation of the inspection system using RGB images of roofs that suffered damage from hail impact. Using the drone to capture images to compose a dataset, the authors obtained results with precision when performing processing by convolutional neural networks to infer the extent of roof damage caused by hail.

In this research, Ottoni, Novo, and Costa [20], and Yeşilmen and Tatar [21], considered the applications of Densely Connected for

Convolutional Networks (Densenet121) and Convolutional Network for Classification and Detection (VGG16) network architectures in processing a roof image bank to classify two classes: roofs with clean gutters and roofs with dirty gutters. One of the research challenges was the definition of hyperparameters, so the authors propose a method for adjusting hyperparameters of convolutional neural networks for the classification of images captured by drones in building construction. The HyperparameterSK (HyperTuningSK) algorithm was developed to create learning rate and optimization rankings, showing the best results for both network architectures used.

In their survey based on a bibliographic review, the authors found a lot of research focused on the acquisition of images by drone and the formation of dataset to identify the state of conservation. In the publications consulted, a technological stage with low experimentation in roof systems with deep learning was found. In the study by Banaszek, Banaszek, and Cellmer [11], the drone acquisition of images of the equipment on the roof is carried out, aiming at the inventory of equipment and technical infrastructure of the buildings, but the application of deep learning was not used.

In particular, no publication was found that reported the detection or classification of equipment arranged on a flat roof, and neither various algorithms nor metrics of performance have been tested, thus justifying the need to advance in these studies involving network architectures capable of differentiating similar objects, in complex environments and with many variables, as is the case with waterproofing flat roofs.

3. Research methodology

3.1. *Experimental environment*

The experiment to detect equipment such as air conditioning condensers on a flat roof with a waterproofing system was carried out in a set of 9 buildings located at the Brazilian Army Headquarters (QGEx). Fig. 1 shows the set of buildings. Depending on the configuration of the buildings, the dimensions of the roofs vary between 200.40 m x 12.50 m and 250.80 m x 14.90 m. Fig. 2 shows the presence of some equipment on the waterproofed slabs of the flat roofs.

3.2. *Parameters and planning for the execution of flights*

To obtain the dataset, the images were captured by the Mavic Pro drone manufactured by DJI - DÀ-JIĀNG INNOVATIONS SCIENCE AND TECHNOLOGY [28], during the execution of three flight plans, planned in the DJI PILOT software [28] (see details in Table 2). The inspections and data acquisition of the condensers were based on navigation by waypoints defined by geographic coordinates, through the Geographic Positioning System (GPS). Fig. 3 presents the drone model used to capture the images. Table 1 lists the parameters required for experiments with the DJI Mavic Pro UAV to obtain images for the annotations of condensers.

For the execution of the three flight plans, it was necessary to establish the flight protocol. The distance adopted between the UAV and the flat roofs was considered the smallest possible to obtain the level of precision and detail in the acquisition of images. To avoid possible collision with the antenna and ladder towers, two flight plans had different altitudes due to these obstacles present in the mapped area. All the flat roofs have the same height of 10.6 m, but the ladder towers have a height of 15.80 m. Only one of the buildings differs from the others, with a height of 8.45 m and the presence of a transmission antenna approximately 15 m high.

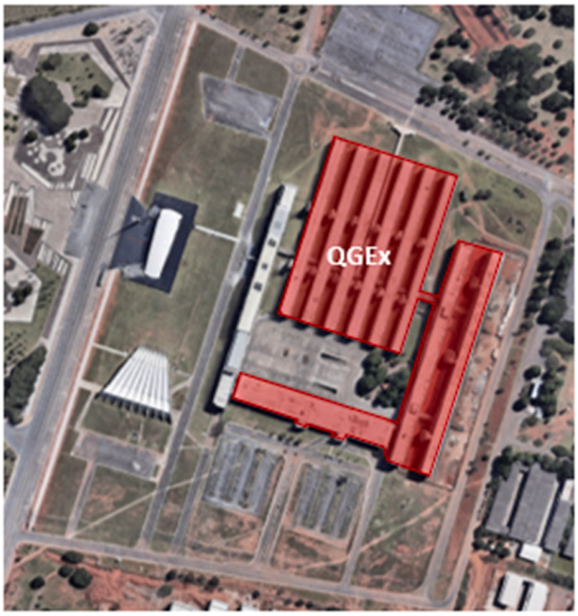


Fig. 1. Location of buildings at the Brazilian Army Headquarters (QGEx) in Brasília-DF.

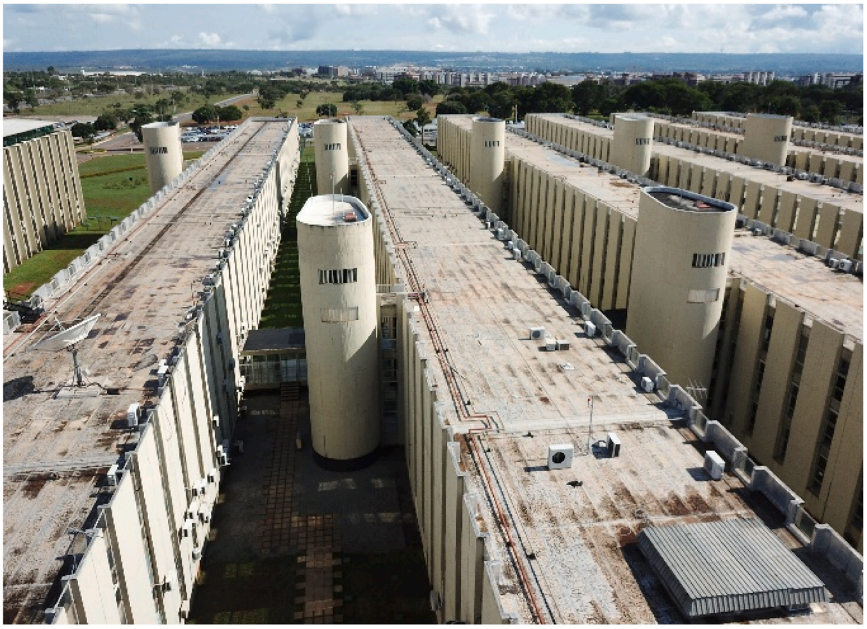


Fig. 2. Equipment arranged on flat roofs.

Fig. 3. Mavic Pro UAV.

Table 1
DJI Mavic Pro UAV camera parameters [28].

Parameters	Value
Sensor	1/2.3″ (CMOS), Effective Pixels: 12.35 M (Total Pixels: 12.71 M)
Lens	FOV 78.8° 26 mm (35 mm equivalent format) f/2.2Distortion < 1.5% focus from 0.5 m to ∞
ISO photo range	100–1600
Electronic shutter speed	8 s – 1/8000 s
Image dimensions	4000 × 3000 pixels
Take-off weight	734

The three flights performed (Fig. 4) to obtain images of the flat roofs were of the cross-flight type (in both directions), with a frontal and lateral overlap rate of 60%. In order to maintain greater uniformity of the images, the researchers sought to avoid variations in positions and configurations. During the flights performed, it was necessary to change the batteries whose autonomy was 27 min.

The first two flights performed on 09/08/2021 at 10:00 am had good weather conditions - a sunny day with a wind speed of 14 km/h. The third flight was performed on 03/07/2022 at 15:30 h and, on that day, the sky was partly cloudy, with a wind speed of 24 km/h. Table 2 organizes information about the flights performed and the weather conditions at the time of the flight, as well as the number of images acquired during each of the flights.

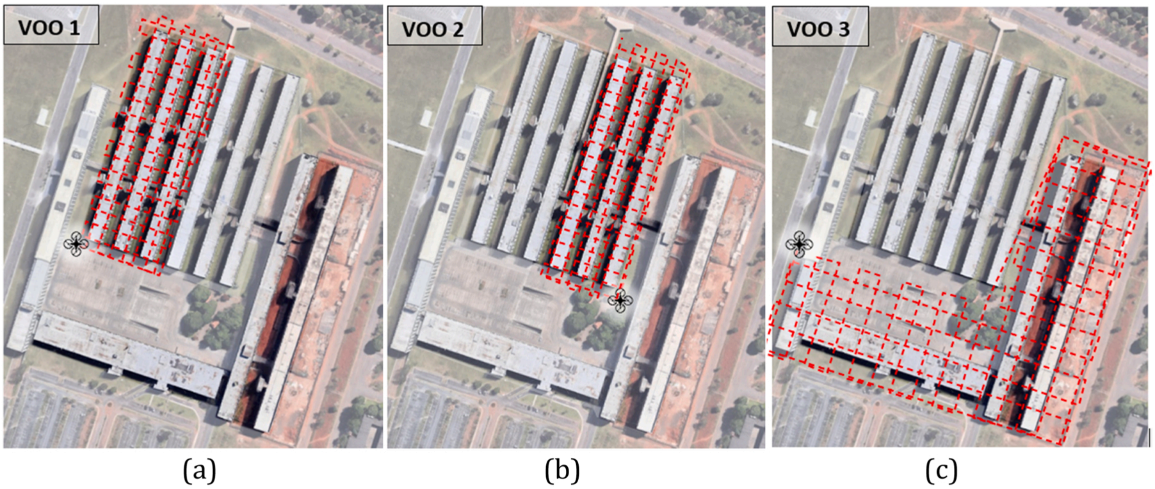


Fig. 4. Flight plans.

Table 2
Performed flights and information about the acquired images.

Site of experimentation	QGEx		
Flight	A	B	Total
Flight date	08/09/21	07/03/22	-
Number of photos collected	428	338	766
Number of flights	2	1	3
Drone	DJI Mavic Pro	DJI Mavic Pro	-
Total flight duration	26 m 01 s	22 m 09 s	48 m 10 s
Quantity of batteries	1	1	2
Height	22 m	35 m	-
Flight type	Automatic 2 directions	Automatic 2 directions	-
Time	10:00 h	15:30 h	-
Covered area	2 ha	5 ha	7 ha

3.3. *Pre-processing the images and building the dataset*

The three flights performed provided 766 images acquired by the Mavic Pro UAV. From this total of captured images, 330 images were selected with high resolution and good luminosity, which contained air conditioning equipment for later annotation in ROBOFLOW [27]. In this stage of pre-processing and construction of the dataset, one of the criteria for selecting the images was to identify the presence of condensers that are positioned directly on the slab, because this is one of the causes of water puddles and drainage obstruction that accelerate the process of degradation of the waterproofing of the roofing system.

The ROBOFLOW platform [27] hosts a set of public data in various formats, including the COCO (Microsoft Common Objects in Context) format which was used for labeling objects in images (Fig. 5 and Fig. 6) and exporting them, providing the necessary tools to convert images into computer vision models. To carry out the training, MMDetection 2.12.0 was used.

The 330 images annotated in ROBOFLOW [27] constituted a dataset whose characteristics are described in Table 3. Fig. 7 shows the distribution of the equipment count per image. On average, the highest count ranged from 2 to 4. Using the heat map shown in Fig. 8, the distribution of the positions in which the condensers appear in the image bank is presented, ranging from 0 to 1372 annotations, considering the pixels contained in each rectangle. Light green pixels are the ones that appear in most annotations.

3.4. *Experimental approach*

Fig. 9 presents the workflow of the research method. First, three flight plans were established, considering the local conditions as barriers to determining the altitude of the flights. Secondly, the flight plan was carried out in loco with the capture of 776 images to form the dataset. In the third stage, 330 images were selected, considering only those that contained the presence of condensers. Then, the selected images were annotated using the ROBOFLOW platform [27]. The annotations were exported and the six network architectures were trained in MMDetection with five-fold cross-validation. Finally, nine metrics were applied to analyze the performance of network architectures through statistical results.

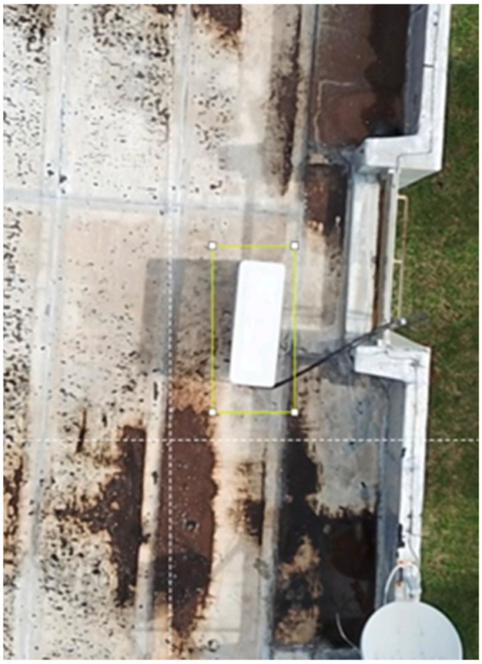


Fig. 5. Check box by ROBOFLOW [27].

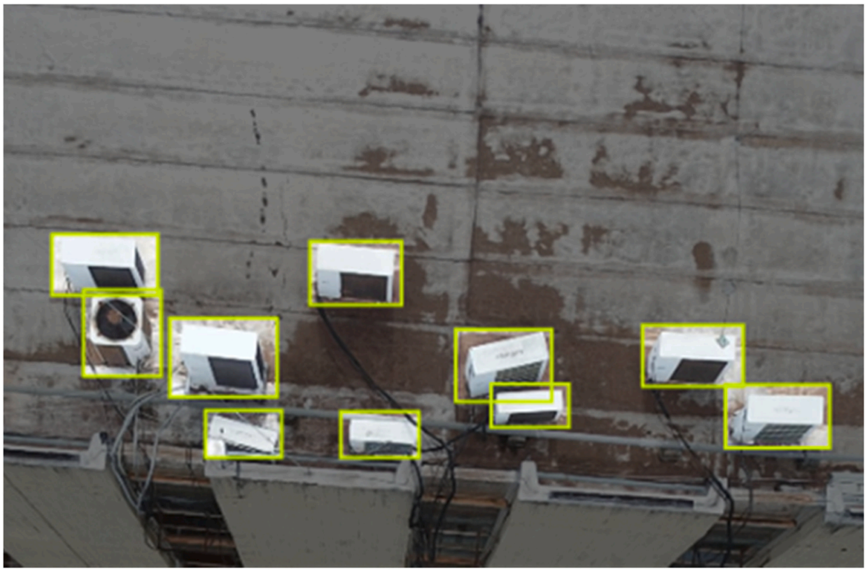


Fig. 6. Annotation of condensers by ROBOFLOW [27].

Table 3
Description of the dataset on the ROBOFLOW platform: Class Balance – Condenser.

Images	Annotations	Average Image Size	Median Image Ratio
330	1372	3.15 MP	2048 × 1680

3.4.1. Deep learning - detection algorithms and hyperparameters

The dataset with 330 annotated images was exported from the ROBOFLOW platform [27] to MMDetection.2.12.0. which is a framework with a toolbox containing a set of object detection methods. For the initialization of the training phase, COCO provided the

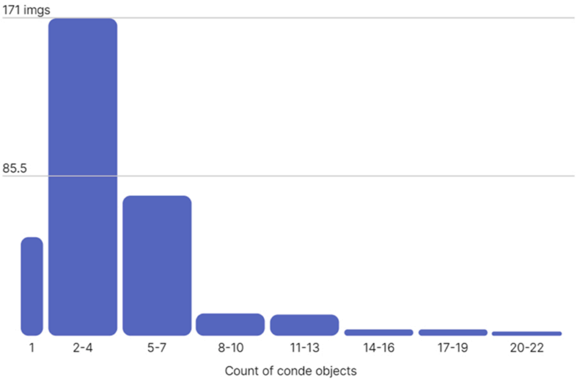


Fig. 7. Histogram of object count by image.

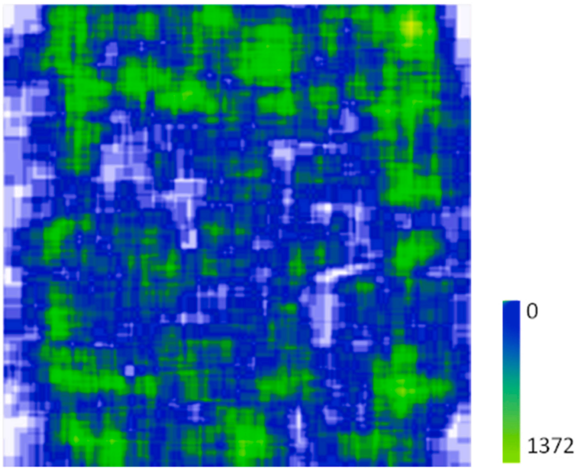


Fig. 8. Annotation Heatmap.

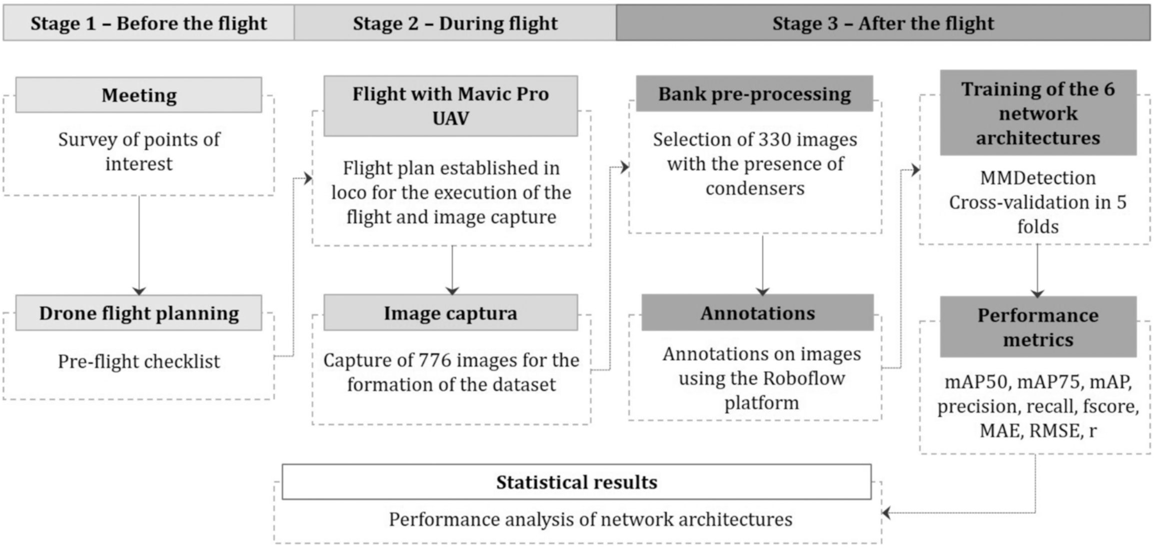

Stage 1 – Before the flight
Stage 2 – During flight
Stage 3 – After the flight
Meeting
Survey of points of interest
Drone flight planning
Pre-flight checklist
Flight with Mavic Pro UAV
Flight plan established in loco for the execution of the flight and image capture
Image captura
Capture of 776 images for the formation of the dataset
Bank pre-processing
Selection of 330 images with the presence of condensers
Annotations
Annotations on images using the Roboflow platform
Training of the 6 network architectures
MMDetection Cross-validation in 5 folds
Performance metrics
mAP50, mAP75, mAP, precision, recall, fscore, MAE, RMSE, r
Statistical results
Performance analysis of network architectures

Fig. 9. Workflow of the research method.

Table 4
Description of the object detection methods.

Symbol Used	Descritption
Faster R-CNN	Region-based convolutional neural network (Faster R-CNN) consists of a two-stage deep learning object detector that uses the Region Proposal Network (RPN) as a region proposal algorithm, and Fast R-CNN as a detector network. Firstly, these networks identify regions of interest and use the region proposal algorithm to generate bounding boxes or locations of possible objects in the image. Then, these regions are transferred to a convolutional neural network that is used to obtain features of these objects followed by a classification layer to predict which class these objects belong to and, finally, a regression layer to make the coordinates of the object bounding box more precise [22].
Retina-Net	Focal Loss (Retina-Net) uses a backbone network and two sub-networks to do specific tasks. The first network, the backbone network is responsible for computing a convolutional feature map over an entire input image. The first subnet performs convolutional object classification on the backbone's output, and the second subnet performs convolutional bounding box regression. Two stages object detectors, Class Imbalance is addressed by a two-stage cascade and sampling heuristics. The proposal stage rapidly narrows down the number of candidate objects' locations to a small number, filtering the background samples. In the second classification stage, sampling heuristics, such as a fixed-foreground-to-background ratio, are performed to maintain a manageable balance between foreground and background [23].
ATSS	Adaptive Training Sample Selection (ATSS) is a method that automatically selects positive and negative samples according to the statical characteristics of the object. It bridges the gap between anchor-based and anchor-free based detectors [24].
VFNet	VarifocalNet (VFNet) is an IoU-aware Dense Object Detector, aimed at accurately ranking a huge number of candidate detections. It consists of a new loss function named Varifocal Loss, for training a dense object detector to predict the IACS (IoU-Aware Classification Score) and a new efficient star-shaped bounding box feature representation for estimating the IACS and refining the coarse bounding boxes [23].
SABL	Side-Aware Boundary Localization (SABL) consists of using a dedicated network branch on each side of the bounding box to localize objects[25].
Fovea	FoveaBox (Fovea) é uma estrutura precisa, flexível e livre de âncoras para detecç ão de objetos. Esta arquitetura de rede aprende diretamente as características do objeto e as coordenadas da caixa delimitadora sem referência de âncora [26].

pre-trained weights to the dataset. The training was carried out in six deep neural networks, namely, Region-based convolutional neural networks (Faster R-CNN), Focal Loss (Retina-Net), Adaptive Training Sample Selection (ATSS), VarifocalNet (Vfnet), Side-Aware Boundary Localization (SABL), and FoveaBox (Fovea). In Table 4 a brief description of these object detection architectures is presented.

To accomplish statistical comparisons of neural network performance, the cross-validation technique was used. This technique basically consists of performing training and testing of the networks using the same dataset as a base, but rearranging different sets for each fold, and randomly selecting to perform new training and testing. For this experiment, five-folds were used for validation.

The hyperparameters used for the experiment are presented in Table 5. The hyperparameters were configured so that the CNNs were trained for 15 epochs, using a learning rate equal to 0.01. The authors consider a value of 0.5 for the classification threshold and 50% for the Intersection over Union (IoU).

3.4.2. Performance Metrics

The six network architectures were compared to verify the performance in detecting equipment condenser-type positioned on flat roofs, using ready-made scripts to generate performance statistics. Nine metrics (Table 6) were used to analyze the performance of the object detection system. For each method tested, the metrics Precision, Recall, fscore, mAP, mAP50, mAP75, MAE, RMSE, and r were calculated.

For this experiment, the code detectors_json_k_dobras[1] was used. This code was developed and made available by the INOVISÃO research group. Statistical analysis was performed using the Rstudio software [29].

The research team used a one-way ANOVA [30] hypothesis test performed at a 5% significance level to check for differences between the six CNN methods by comparing their respective performance metric values. The condition of homogeneity of the groups and normal distribution were checked, both of which are necessary to apply ANOVA and Tukey's post-hoc test [32]. In addition, the non-parametric Kruskal-Wallis test was applied [31]. This decision was based on the fact that the cross-validation was performed five times and this number of results may not be representative enough for an ANOVA test. Including, the data were also analyzed using descriptive statistics, boxplot, loss convergence curve, and Pearson Correlation to analyze whether there is a relationship between the two variables (measured and predicted values).

4. Results

Table 7 lists the results of the nine metrics used to evaluate the performance of each of the tested methods. The values presented are the arithmetic means resulting from the means of the five folds.

Comparing the respective performance metric values of the six CNN methods, the application of one-way ANOVA did not show us that there are statistical differences between the means of mAP, mAP50, mAP75, Precision, Recall, fscore, MAE, and RMSE. In all cases, $p > 0.05$ was obtained. Thus, the Tukey's post-hoc test was not applied. Fig. 10 shows the results of ANOVA, means, standard deviation, and confidence interval for the metrics and respective methods.

In Fig. 11, the boxplots for the distribution of the metric values were obtained during processing for each of the methods. Except for the Vfnet method, the other five methods showed little variability in the distribution of the measured values. Vfnet shows a very wide dispersion in all metrics when compared to the other methods, except for the r metric (Pearson Correlation).

[1] Available: http://git.inovisao.ucdb.br/inovisao/detectores_json_k_dobras [29]

Table 5
Description of the hyperparameters used for the experiment.

Hyperparameters	Description
Epochs	The epochs for an artificial intelligence system refer to the number of times that the algorithm will perform training, tests, and validations with the entire available dataset, adjusting the iteration weights based on the results of each executed epoch.
Classifier threshold	Defines the threshold that determines whether or not the detected object is the object to be detected.
Intersection Over Union (IoU)	Determines the percentage that the annotation performed by the algorithm must intersect with the manual annotation to be considered a detection.
Learning Rate	Determines the size of step that each learning iteration should have relative to the whole, as that step moves to the minimum of a loss function.

Table 6
Description of performance metrics used for evaluation.

Metric	Symbol Used	Descritption
Precision	P	It is considered the fraction of the trues positives, correct prediction from the total amount of relevant results, i.e., the sum of TP and FP. **P = TP/(TP+FP)**
Recall	R	It is defined as the fraction of real positive cases that are related correctly, i.e., the quantity of TP from the total amount of TP and FN. In our experiment, how the algorithm works with one-class detection problems, the FN is defined as any detection that doesn't intercept the box area done manually. In this case, the FN is any detection that has IoU equal to zero. **R=TP/(TP+FN)**
F1score	F1	It is the harmonic mean of precision and recall. **F1 = 2 x (TP x FP)/(TP+FP)**
Mean Average Precision	mAP	The mAP value represents the area under the Precision-Recall curve, which connects the pairs (Px; Ry) in a range from 0 to 1.
	mAP50 mAP75	AP50 and AP75 refer to the value fixed for the Intersection Over Union - IoU threshold, respectively, at 50% or 75%. In this case, mAP50 and mAP75 represent the mean of values obtained for each fold.
Mean Absolute Error	MAE	It is the magnitude of the difference between an expected and predicted value. As this result can be positive or negative, it is suggested to use the MAE math function for the average of absolute error values always returns positive values.
Root Mean Square Error	*RMSE*	It is the standard deviation of the differences between predicted values and observed values.
Pearson Correlation	*r*	The coefficient shows the relationship between two variables. When r is equal to zero, in this case, there is no correlation between the variables. The closer the coefficient is to 1 or − 1, the stronger the relationship. If the correlation is negative, the variable is influencing the other in the opposite direction. If positive, both variables are increasing or decreasing in the same direction.

*True Positive-TP; False Positive-FP; True Negative-TN; False Negative-FN.

Table 7
Methods used and performance metrics values.

	mAP	mAP 50	mAP 75	Precision	Recall	Fscore	MAE	RMSE	r
Faster R-CNN	0.654	0.944	0.814	0.927	**0.934**	**0.930**	**0.439**	**0.914**	**0.959**
Retina-Net	0.513	0.750	0.628	0.713	0.730	0.721	1.324	2.024	0.746
ATSS	**0.665**	0.943	**0.825**	**0.959**	0.835	0.893	0.706	1.527	0.907
Vfnet	0.399	0.569	0.492	0.537	0.549	0.543	1.976	2.779	0.699
SABL	0.659	**0.951**	0.816	0.920	0.893	0.906	0.549	1.177	0.934
Fovea	0.642	0.935	0.796	0.919	0.889	0.903	0.597	1.208	0.931

Label: The best performance The second-best performance The third best performance

Pearson Correlation is shown in Fig. 12 for each method. The authors analyzed the r-metric ($r > 0.72$) which shows that the relationship between the measured variables and predicted values is a strong relationship and positive. Retina-Net demonstrates a better correlation ($r = 0.87$) followed by Fovea ($r = 0.857$), and Faster R-CNN ($r = 0.83$), respectively. The graphs in Fig. 12 show the regression lines that were constructed using all the 5-fold test images. This is the reason that explains the difference in the value of the RMSE, MAE, and r metrics presented in the graphs of Fig. 12, in relation to the values of the metrics calculated by fold, shown in Table 7.

When applying ANOVA for the r-metric, the evidence shows that there are significant differences ($p = 0.0219$) between the group mean when compared to the means of each method. In this case, considering that $p < 0.05$, the post hoc Tukey test was applied for the multiple comparisons of means, for a 95% confidence level. When comparing the methods with each other, statistical differences were found between the ATSS and Faster R-CNN methods, for $p = 0.006$.

When the r-metric is analyzed, the results measured in the five folds show less variability among themselves. Stands out that when ANOVA was applied, it was not possible to state that there are statistical differences between the groups. However, when applying the non-parametric Kruskal-Wallis test for the metric r, $H (5, N = 30) = 11.79$ and $p = 0.0377$ were obtained, which explains that there

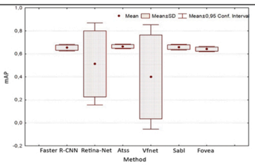
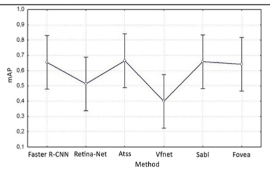
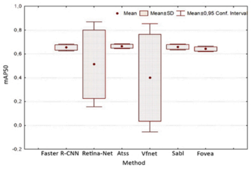
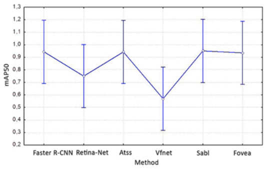
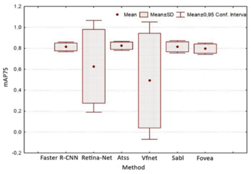
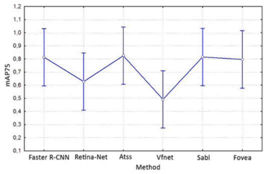
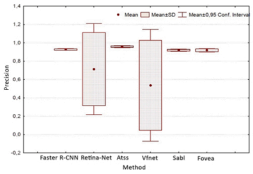
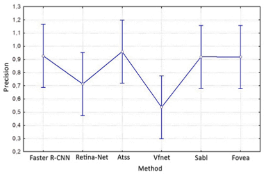
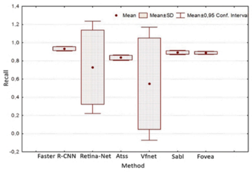
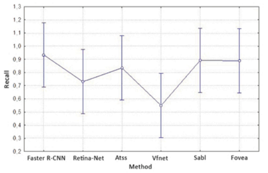
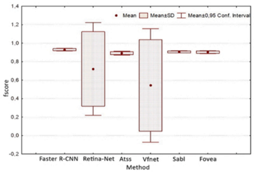
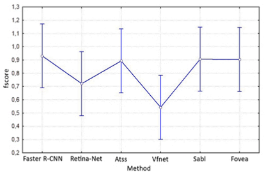
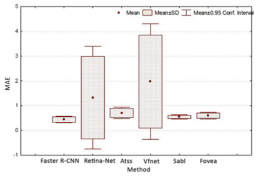
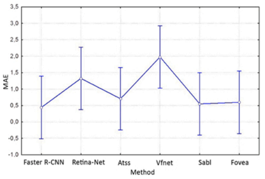
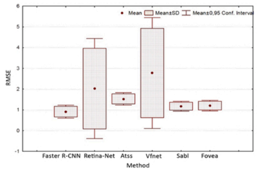
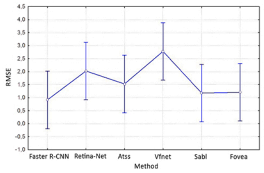
Metrics	Methods		
mAP			Least Squares Means (LS Means) Current effect: F(5, 24) =1,6386, p=0,18807; Confidence intervals=0,95
mAP50			Least Squares Means (LS Means) Current effect: F(5, 24) =1,6626, p=0,18203; Confidence intervals=0,95
mAP75			Least Squares Means (LS Means) Current effect: F(5, 24) =1,6626, p=0,18203; Confidence intervals=0,95
Precision			Least Squares Means (LS Means) Current effect: F(5, 24) =2,1167 p=0,09817; Confidence intervals=0,95
Recall			Least Squares Means (LS Means) Current effect: F(5, 24) =1,4759 p=0,23449; Confidence intervals=0,95
Fscore			Least Squares Means (LS Means) Current effect: F(5, 24) =1,7338 p=0,16521; Confidence intervals=0,95
MAE			Least Squares Means (LS Means) Current effect: F(5, 24) =1,6906 p=0,17523; Confidence intervals=0,95
RMSE			Least Squares Means (LS Means) Current effect: F(5, 24) =1,6583 p=0,18310; Confidence intervals=0,95

Fig. 10. Graphs showing the metrics values of each method and respective means, standard deviation, and confidence interval.

are differences between the groups. When performing multiple comparisons, evidence shows that there are statistically significant differences ($p = 0.037$) for the adopted significance level ($p < 0.05$) between the medians of Faster R-CNN and ATSS, thus explaining the boxplot in Fig. 12 for the r-metric.

The trained methods' learning curves are presented in Fig. 13. In general, it was noticed that the learning process occurred in all

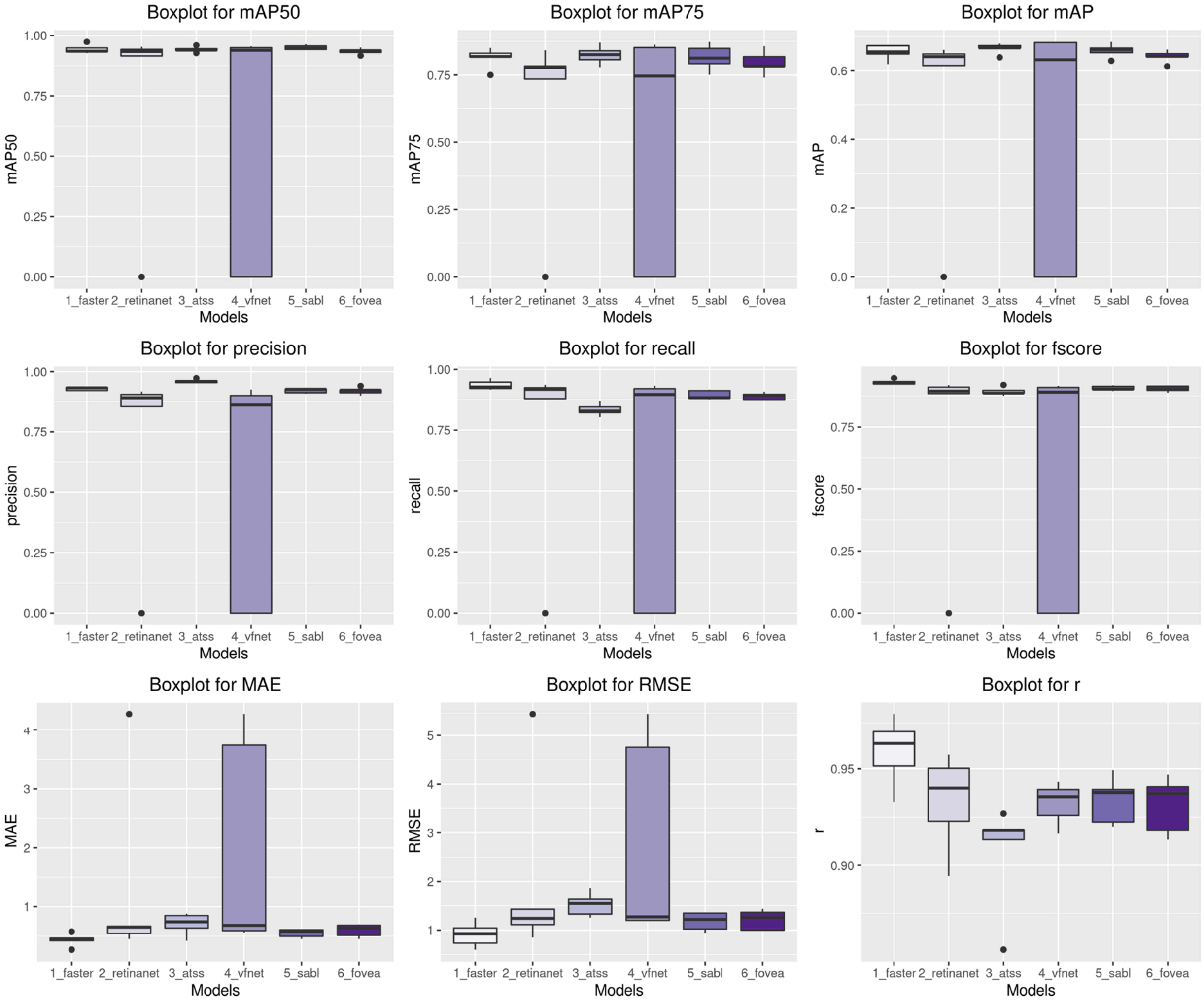


Fig. 11. Boxplot for the metric values measured in each experiment.

methods. Firstly, it is observed the loss decrease in the curves during the validation process, according to the epochs. This means that the algorithms were able to generalize the problem using the data during the training step.

In addition, the evolution of the loss curve during training shows that the Faster R-CNN and ATSS methods showed stabilization from the tenth epoch. Nonetheless, the Retina-Net, Vfnet, SABL e Fovea methods showed instability in the learning step, with an eventual tendency for errors.

The loss curves (Fig. 13) show the importance to adopt many epochs to train the algorithms when it is possible to check if occurred learning stabilization process in the network and the non-occurrence of overfitting. If the error is not decreasing, it can be deduced that the algorithm, in this situation, can't generalize due to a small validation or training data set, or indicates an overfitting problem. However, the curve behavior can indicate that the adopted learning rate was high, which makes it difficult to learn better for some methods.

As a result of the non-parametric Kruskal-Wallis test, it is not possible to state that the metrics mAP50, mAP75, mAP, MAE, and RMSE presented statistically significant differences between the evaluated methods. The metric r ($p = 0.038$) showed a statistically significant difference only between the Faster R-CNN and ATSS methods. However, the Kruskal-Wallis test showed statistically significant differences between the methods, when analyzing the metrics Precision ($p = 0.0009$), Recall ($p = 0.0193$), and fscore ($p = 0.0142$) shown in Table 8 below. In bold, results are presented where $p < 0.05$, i.e., when comparisons between two methods resulted in statistically significant differences. Table 8 shows that the multiple comparisons performed by the Kruskal-Wallis test show significant differences for Precision, between the ATSS _ Retina-Net and ATSS _ Vfnet methods. For the Recall metric, differences were observed between ATSS _ Faster R-CNN. Finally, fscore showed differences between ATSS _ Faster R-CNN and Vfnet _ Faster R-CNN.

ATSS achieved the best Precision (0.959), followed by Faster R-CNN (0.927) and SABL (0.920). Evidence showed that it is not possible to indicate whether there are statistical differences between these CNNs, but that Retina-Net and Vfnet had the worst performances. Faster R-CNN had the best Recall (0.934), followed by SABL (0.893) and Fovea (0.889).

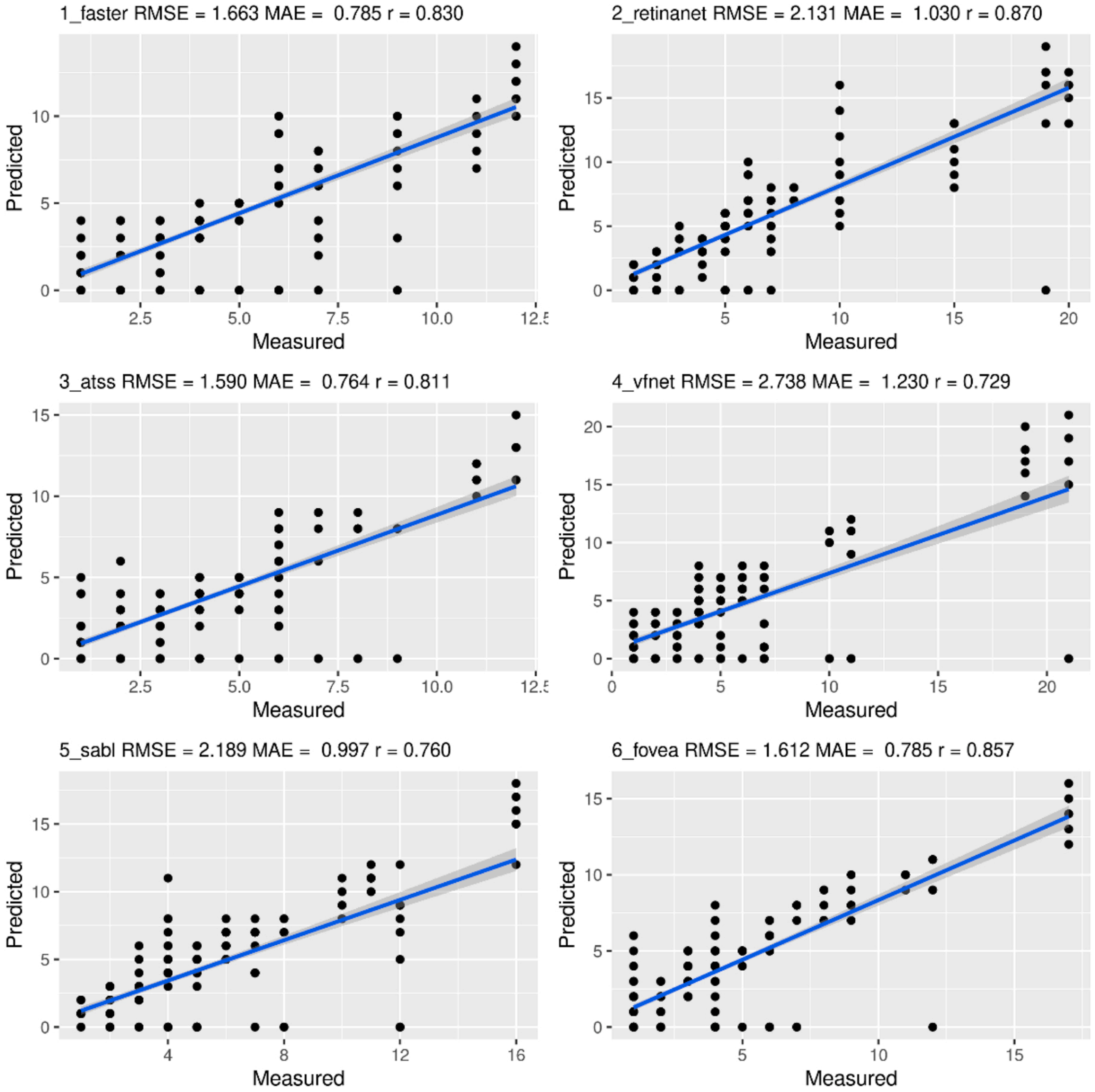


Fig. 12. Results of Pearson Correlation: relationship between the measured and predicted values.

In turn, it is not possible to state that there are statistical differences between these CNNs, but that ATSS (0.835) differs from Faster R-CNN and that Retina-Net and Vfnet had the worst performances. Faster R-CNN had the best Fscore (0.930), followed by SABL (0.906) and Fovea (0.903).

Overall, analyzing the results of Table 7 and Table 8 together, it appears that all tested CNN showed good performance, with emphasis on Faster R-CNN followed by SABL, Fovea, and ATSS. The neural networks Retina-Net and Vfnet were those that presented the lowest values in the nine tested metrics.

5. Discussion

In general, even though convolutional neural networks perform well, the interest in this discussion is to verify why the errors occurred since the dataset was built for this study and had never been tested before. In fact, the problem addressed in this article is still a research gap, as no publications were found in a bibliographic review reporting object detection in building roofs, using neural networks. Therefore, it is primordial to map the errors and to find the causes, from the annotation.

After the CNN training period, with 1372 annotations in a dataset of 330 images, the condensers that were correctly identified and the mistakes made both were possible to analyze. When comparing the manual and automatic annotations, Faster R-CNN was the network that made more correct annotations. On the other hand, the most recurrent error was the duplication of automatic annotation,

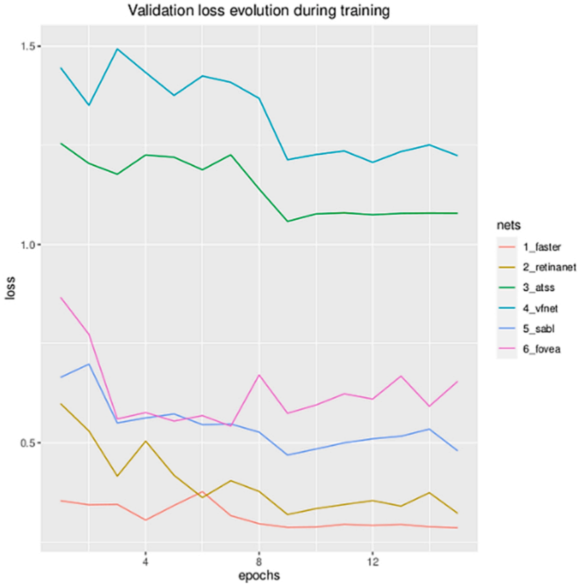


Fig. 13. Loss Curves: validation loss evolution during training.

Table 8
Results of non-parametric Kruskal-Wallis test: $p < 0.05$.

Multiple Comparisons	Methods	Faster R-CNN	Retina-Net	ATSS	Vfnet	SABL	Fovea
Independent variable: Precision	**Faster R-CNN**		0.338	1.000	0.354	1.000	1.000
Kruskal-Wallis test: H (5, N = 30) = 20.862;	**Retina-Net**	0.338		**0.002**	1.000	1.000	1.000
p = 0.0009	**ATSS**	1.000	**0.002**		**0.002**	0.558	0.354
	Vfnet	0.354	1.000	**0.002**		1.000	1.000
	SABL	1.000	1.000	0.558	1.000		1.000
	Fovea	1.000	1.000	0.354	1.000	1.000	
Independent variable: Recall	**Faster R-CNN**		1.000	**0.005**	0.488	0.889	0.467
Kruskal-Wallis test: H (5, N = 30) = 13.470;	**Retina-Net**	1.000		0.510	1.000	1.000	1.000
p = 0.0193	**ATSS**	**0.005**	0.510		1.000	1.000	1.000
	Vfnet	0.488	1.000	1.000		1.000	1.000
	SABL	0.889	1.000	1.000	1.000		1.000
	Fovea	0.467	1.000	1.000	1.000	1.000	
Independent variable: fscore	**Faster R-CNN**		0.060	**0.030**	**0.020**	0.636	0.338
Kruskal-Wallis test: H (5, N = 30) = 14.232;	**Retina-Net**	0.060		1.000	1.000	1.000	1.000
p = 0.0142	**ATSS**	**0.030**	1.000		1.000	1.000	1.000
	Vfnet	**0.020**	1.000	1.000		1.000	1.000
	SABL	0.636	1.000	1.000	1.000		1.000
	Fovea	0.338	1.000	1.000	1.000	1.000	

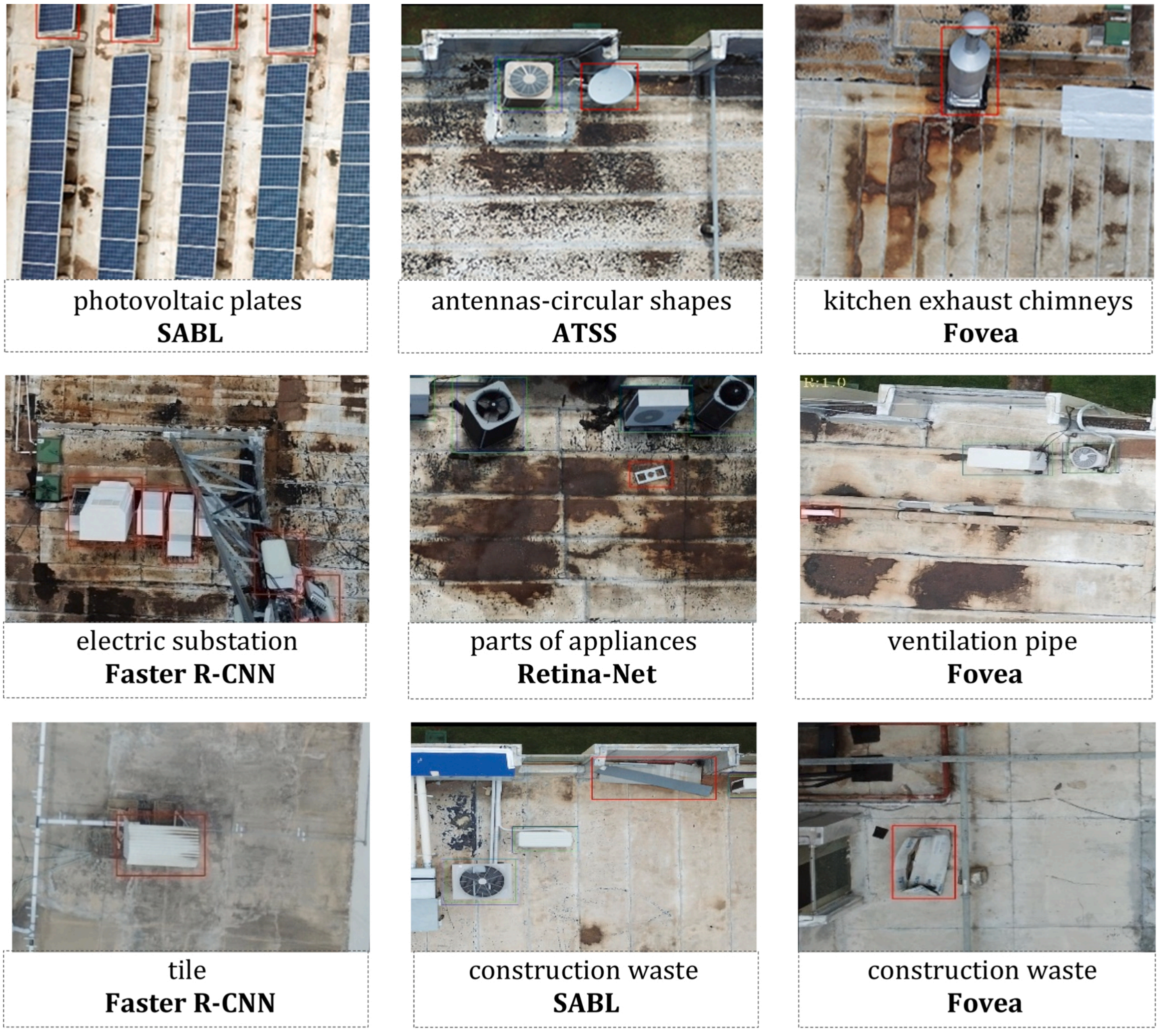
mainly by Retina-Net and SABL networks. Another important error identified was the loss of image annotations made on the ROBOFLOW platform [27], which were not carried over to the neural networks trained in MMDetection.

The detection of similar objects was also one of the most common types of errors made by all networks (Fig. 14). In the case of the flat roof under study, for the neural networks tested, the similar objects are the antennas with circular shapes, the kitchen exhaust chimneys, the rectangular substations of the communication antennas in white color, and parts of deactivated appliances, among others more sporadic situations such as construction waste deposited on the slab, photovoltaic plates, tiles, and ventilation pipes. Incorrectly detected objects have rectangular or circular shapes, with characteristics similar to condensers.

Many objects were identified even though they were not annotated. For example, the automatic annotations made on the condensers on the facades of the buildings (Fig. 15) and the white cars (ambulance) on the sidewalk on the ground floor. This kind of mistake was mainly made by Fovea.

Fig. 16 shows the pooling of water that occurs near the drains, caused by the equipment installed there. The image captured by the UAS has good resolution, which illustrates how these types of equipment positioned in inappropriate locations can accelerate the degradation process of the waterproofing membrane, due to the presence of moisture and mold formation, impacting the performance of construction systems.

The maintenance performed on the waterproofing blanket caused a color difference and, therefore, was detected. These waterproofing repairs stand out for their rectangular shape and lighter coloring in relation to the darker color of the background. Fig. 17



photovoltaic plates
SABL

antennas-circular shapes
ATSS

kitchen exhaust chimneys
Fovea

electric substation
Faster R-CNN

parts of appliances
Retina-Net

ventilation pipe
Fovea

tile
Faster R-CNN

construction waste
SABL

construction waste
Fovea

Fig. 14. Types of errors made by neural networks when detecting similar objects.

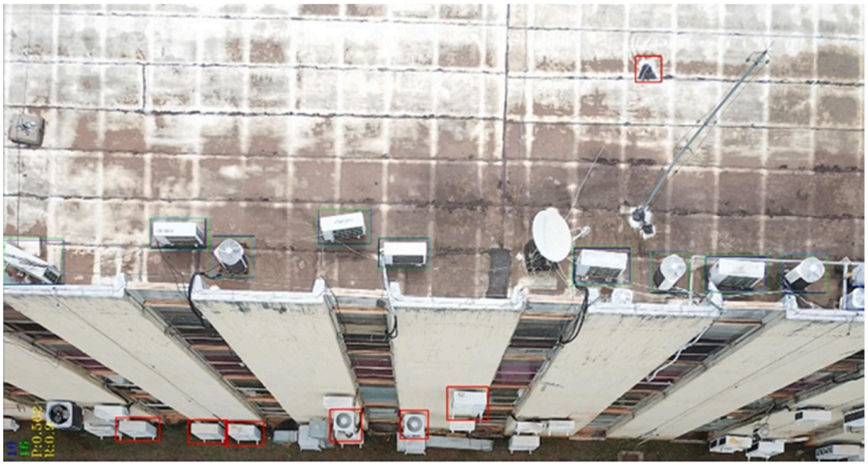


Fig. 15. Detection errors of condensers on the facade by the Fovea network.

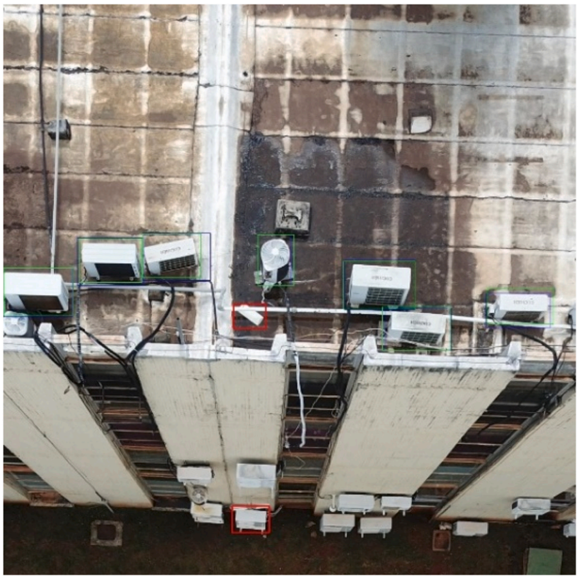


Fig. 16. Pooling of water in the region of the condensers.

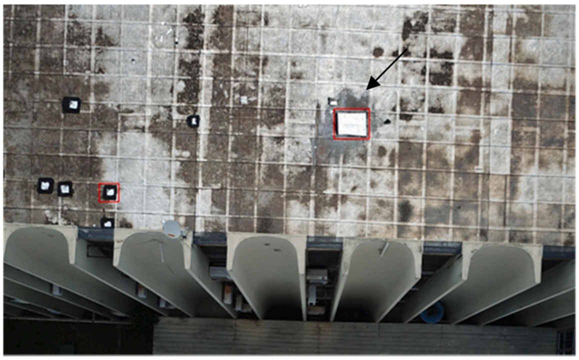


Fig. 17. Repair carried out on waterproofing blanket.

shows the detection of waterproofing repairs identified as an error in Faster R-CNN, Retina-Net, Vfnet, SABL, and Fovea networks. Condensers with two or more exhaust fans or even condensers very close to each other as a result of the duplication of the object by automatic annotation, caused the incorrect identification of the bounding box that was displaced from the manually annotated parameter, influencing a low IoU value.

The neural networks had difficulties in detecting the condensers in the images in which the capture angle in relation to the object was very close to the right angle (90°) and also in images with a pigmented background. These discolored areas may have interfered with the image homogeneity, causing image noise that can influence the correct convolution of pixels to identify the format and pattern of the condensers. The images obtained in the higher altitude flight presented a low resolution, because of the size of the condensers. Consequently, the algorithms had difficulties identifying them. The images that neural networks failed to identify correctly can affect the Recall metric.

Retina-Net was the network that had the worst detection performance, presenting the highest number of errors (Fig. 18). The errors found in the Retina-Net network were concentrated in the detection of condensers that had already been noted, in the detection of the condensers present in the facades, and in the detection of substations with similar shapes and colors of condensers.

When the performance is analyzed based on theory, it is possible to discuss the fact that Retina-Net is a one-stage detection network but seeks to achieve two-stage performance while keeping computational complexity down. In the experiment performed, this assumption was not confirmed, since its performance did not surpass other network architectures, including Faster R-CNN, which is a two-stage network for detection.

On the other hand, ATSS is a method that bridges the gap between anchor-based and anchor-free detectors. When analyzing the good performance of ATSS for a value of Precision = 0.959 (highest among all tested networks), the advantage of a specialized detection method capable of automatically selecting positive and negative samples, according to the static characteristics of the object,

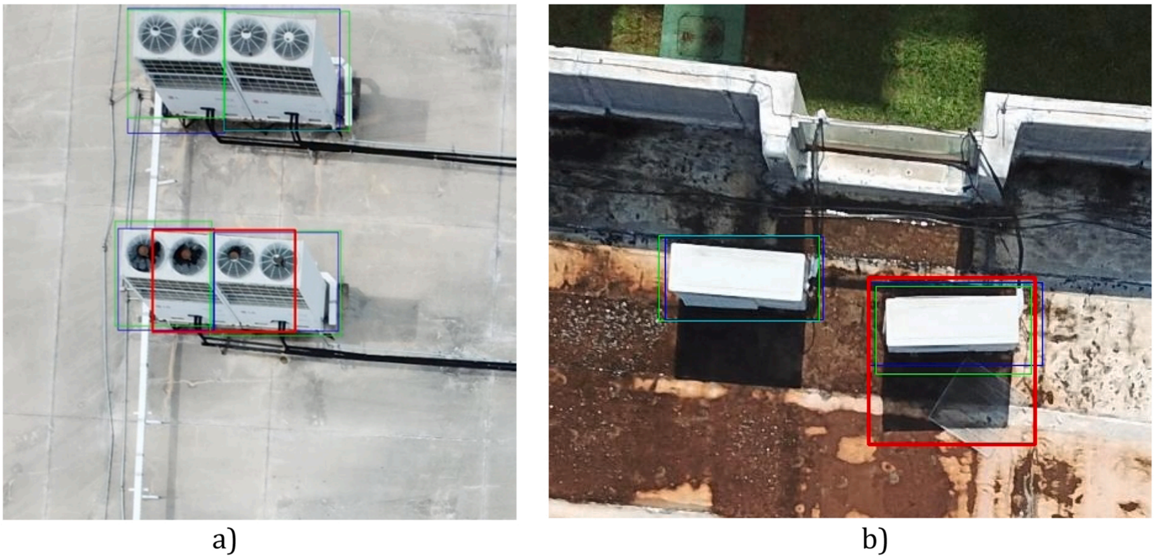


Fig. 18. Error in detecting condensers on the Retina-Net network.

is discussed.

6. Conclusion

The six methods tested in the experiment were evaluated by nine performance metrics and the results show that the dataset consists of a set of images capable of being measured by deep learning, thus contributing to the advancement of automated processes in real-time to decision-making for maintenance planning.

Due to the pioneering character of the study in flat roofs, the authors highlight the importance of the experiment that tested six convolutional neural networks, evaluated in nine metrics, considering both the Analysis of Variance (ANOVA one way) and the non-parametric Kruskal-Wallis test. The mistakes made and mapped, as well as the identification of the causes, are impacting factors in the decision making regarding the selection of CNN and the acquisition of images to compose the dataset. Among the evaluated methods, the best results were provided by Faster R-CNN, ATSS, and SABL.

Equipment such as condensers can be detected using computer vision. The quality of the dataset positively influenced the good performance of the neural networks, although the types of errors made in the detection alert to the diversity of causes related to the inherent characteristics of flat roofs. Evaluating the errors made by the networks, when considering geometric shapes and colors similar to the condensers, proximity between objects, and location of objects on facades and sidewalks that are different plans from those predicted in the annotations, it is verified the need to improve algorithms and architectures of the network with the capacity to differentiate similar objects in complex environments, with many variables.

It is concluded that the problem addressed in the experiment cannot be considered trivial, nor can studies on flat roofs be treated from the perspective of methods and metrics established in other areas. In the continuity and improvement of automated processes, one must consider the non-homogeneity of materiality in the construction of buildings, their specificities, and the variability found in their systems and components. Mainly and finally, the data and results of this experiment can be used as a standard for evaluating roof equipment detection algorithms.

7. Future work

During the experiments on flat roofs, due to the flight angle of the drone, part of the building's facades was also captured in the images. The condensers located on these facades had not been manually annotated, because the objective of the research was to detect the equipment on the flat roofs (horizontal plane). However, some neural networks automatically detected the condensers present on the facades (vertical plane). For future work, it is interesting to consider another type of network for pre-processing (or post-processing) with segmentation techniques to automatically exclude regions that do not correspond to the roofs, disregarding the façade regions. Thus, future research on methods of semantic segmentation, detection, and classification of existing equipment in flat roofs becomes potentially innovative for the management of maintenance in the area of Architecture, Engineering, Construction, and Operation.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to

influence the work reported in this paper.

Data Availability

Data will be made available on request.

Acknowledgements

This work has received financial support from the Dom Bosco Catholic University and the Foundation for the Support and Development of Education, Science and Technology from the State of Mato Grosso do Sul, FUNDECT. We would also like to thank NVIDIA for providing the Titan X GPUs used in the experiments. Some of the authors have been awarded with Scholarships from the University of Brasília, Brazilian National Council of Technological and Scientific Development, CNPq and the Coordination for the Improvement of Higher Education Personnel, PROAP/CAPES.

References

[1] G.T. Ferraz, J. Brito, V.P. Freitas, J.D. Silvestre, State-of-the-Art Review of Building Inspection Systems, 04016018-04010188, in: Journal of Performance of Constructed Facilities, [s.l.], v. 30, American Society of Civil Engineers (ASCE), 2016, https://doi.org/10.1061/(asce)cf.1943-5509.0000839.

[2] M. Bown, K. Miller, The use of unmanned aerial vehicles for sloped roof inspections – considerations and constraints, J. Facil. Manag. Educ. Res. Vol. 2 (n. 1) (2018) 12–18.

[3] M. Hezaveh, C. Kanan, C. Salvaggio, Roof Damage Assessment using Deep Learning, in: 2017 Ieee Applied Imagery Pattern Recognition Workshop (Aipr), [S.L.], v. 1, IEEE,, 2017, pp. 1–6, https://doi.org/10.1109/aipr.2017.8457946.

[4] H. Kim, J. Lee, E. Ahn, S. Cho, M. Shin, S. Sim, Concrete crack identification using a UAV incorporating hybrid image processing, Sensors 17 (9) (2017) 2052, https://doi.org/10.3390/s17092052.

[5] B. Silveira, R. Melo, D.B. Costa, Using UAS for Roofs Structure Inspections at Post-occupational Residential Buildings. Lecture Notes in Civil Engineering, Springer International Publishing, 2020, pp. 1055–1068, https://doi.org/10.1007/978-3-030-51295-8_73.

[6] T. Rakha, A. Liberty, A. Gorodetsky, B. Kakillioglu, S. Velipasalar, Heat Mapping Drones: an autonomous computer-vision-based procedure for building envelope inspection using unmanned aerial systems (uas), in: Technology|Architecture + Design, [S.L.], v. 2, Informa UK Limited, 2018, pp. 30–44, https://doi.org/10.1080/24751448.2018.1420963.

[7] D. Roca, S. Lagüela, L. Díaz-Vilariño, J. Armesto, P. Arias, Low-cost aerial unit for outdoor inspection of building façades, in: Automation In Construction, [S.L.], v. 36, Elsevier BV, 2013, pp. 128–135, https://doi.org/10.1016/j.autcon.2013.08.020.

[8] T.D. Akinosho, L.O. Oyedele, M. Bilal, A.O. Ajayi, M.D. Delgado, O.O. Akinade, et al., Deep learning in the construction industry: A review of present status and future innovations, J. Build. Eng. 32 (2020) 1–14.

[9] C. Koch, K. Doycheva, V. Kasireddy, B. Akinci, P. Fieguth, Corrigendum to "a review on computer vision-based defect detection and condition assessment of concrete and asphalt civil infrastructure" [Advanced Engineering Informatics 29(2), 2015, pp. 196–210, Adv. Eng. Inform. 30 (2016) 208–210, https://doi.org/10.1016/j.aei.2016.03.002.

[10] F. Nex, F. Remondino, Preface: latest developments, methodologies, and applications based on uav platforms, in: Drones, [S.L.], v. 3, MDPI AG,, 2019, pp. 1–3, https://doi.org/10.3390/drones3010026.

[11] A. Banaszek, S. Banaszek, A. Cellmer, Possibilities of use of UAVS for technical inspection of buildings and constructions, in: Iop Conference Series: Earth and Environmental Science, [S.L.], v. 95, IOP Publishing, 2017, 032001, https://doi.org/10.1088/1755-1315/95/3/032001.

[12] J. Zhu, J. Zhong, T. Ma, X. Huang, W. Zhang, Y. Zhou, Pavement distress detection using convolutional neural networks with images captured via UAV, Autom. Constr. 133 (2022) 1–11, https://doi.org/10.1016/j.autcon.2021.103991.

[13] L.B. Staffa, L.S. Sá, M.I.S.C. Lima, D.B. Costa, Uso de técnicas de processamento de imagem para inspeção de estruturas de telhados de edificações para fins de assistência técnica. ENCONTRO NACIONAL DE TECNOLOGIA DO AMBIENTE CONSTRUÍDO, ANTAC, Porto Alegre, Anais, 2020, pp. 1–8.

[14] S. Taoufiq, B. Nagy, C. Benedek, Hierarchy Net: hierarchical cnn-based urban building classification, in: Remote Sensing, [S.L.], v. 12, MDPI AG, 2020, p. 3794, https://doi.org/10.3390/rs12223794.

[15] J. Moore, H. Tadinada, K. Kirsche, J. Perry, F. Remen, Z.T.H. Tse, Facility inspection using UAVs: a case study in the University of Georgia campus, Int. J. Remote Sens. v. 39 (n. 21) (2018) 7189–7200, https://doi.org/10.1080/01431161.2018.1515510.

[16] L.M.A. Santos, V.A.G. Zanoni, Deep learning e suas possibilidades de aplicação no patrimônio cultural edificado, in: Patrimônio 4.0: conectando dimensões e realidade, 2022, Goiânia. Anais patrimônio 4.0: conectando dimensões da realidade, v. 1, LaSUS FAU, Brasília, 2022, pp. 1–492.

[17] L.S. Silva, V.A.G. Zanoni, V.C. Pazos, L.M.A. Santos, T.R.P. Jucá, Fotogrametria com imagens adquiridas com drones: do plano de voo ao modelo 3D [livro eletrônico] Brasília, DF: LaSUS FAU: Editora Universidade de Brasília, 2022, https://doi.org/10.29327/563260.

[18] Y. Lecun, Y. Bengio, G. Hinton, Deep learning, in: Nature, [S.L.], v. 521, Springer Science and Business Media LLC, 2015, pp. 436–444, https://doi.org/10.1038/nature14539.

[19] D. Yudin, A. Naumov, A. Dolzhenko, E. Patrakova, Software for roof defects recognition on aerial photographs, in: Journal Of Physics: Conference Series, [S.L.], v. 1015, IOP Publishing, 2018, pp. 1–10, https://doi.org/10.1088/1742-6596/1015/3/032152.

[20] A.L.C. Ottoni, M.S. Novo, D.B. Costa, Hyperparameter tuning of convolutional neural networks for building construction image classification, Vis. Comput. (2022) 1–15, https://doi.org/10.1007/s00371-021-02350-9.

[21] S. YEŞILMEN, B. TATAR, Efficiency of convolutional neural networks (CNN) based image classification for monitoring construction related activities: a case study on aggregate mining for concrete production (dez.), in: Case Studies In Construction Materials, [S.L.], v. 17, Elsevier BV,, 2022, pp. 1–11, https://doi.org/10.1016/j.cscm.2022.e01372.

[22] S. Ren, K. He, R. Girshick, J. Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. arXiv:1506.01497v3. https://doi.org/10.48550/arXiv.1506.01497. Advances in Neural Information Processing Systems 28 (NIPS 2015). Proceeding ISBN: 9781510825024.

[23] T. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal Loss for Dense Object Detection (ArXiv), Comput. Vis. Pattern Recognit., [S. L.] v. 2 (n. 1) (2018) 1–10, https://doi.org/10.48550/ARXIV.1708.02002.

[24] S. Zhang, C. Chi, Y. Yao, Z. Lei, S.Z. Li, Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection (ArXiv), Comput. Vis. Pattern Recognit., [S. L.] v. 1 (n. 1) (2019) 1–10, https://doi.org/10.48550/ARXIV.1912.02424.

[25] J. Wang, W. Zhang, Y. Cao, K. Chen, J. Pang, T. Gong, J. Shi, C.C. Loy, D. Lin, Side-aware boundary localization for more precise object detection (ArXiv), Comput. Vis. Pattern Recognit., [S. L.] v. 1 (n. 1) (2020) 1–21, https://doi.org/10.48550/ARXIV.1912.04260.

[26] T. Kong, F. Sun, H. Liu, Y. Jiang, L. Li, J. Shi, FoveaBox: beyound anchor-based object detection, in: Ieee Transactions On Image Processing, [S.L.], v. 29, IEEE,, 2020, pp. 7389–7398, https://doi.org/10.1109/tip.2020.3002345.

[27] ROBOFLOW. Give your software the sense of sight. 2022. Available at: https://roboflow.com/. Accessed on: January 8, 2023.

[28] DÀ-JIĀNG INNOVATIONS SCIENCE AND TECHNOLOGY - DJI. DJI Fly. 2023. Available at: https://www.dji.com/br/downloads/djiapp/dji-fly. Accessed on: January 8, 2023.

[29] RSTUDIO, Statisticall software. Version 4.2.2. Posit Software, PBC formely Rstudio, PBC. Accessed on: August 1, 2022.

[30] S. Lars, W. Svante, Analysis of Variance (ANOVA), in: Chemetrics and Inteligenct Laboratory Systems, v. 6, Elsevier Publisher,, 1989, p. 4, https://doi.org/10.1016/0169-7439(89)80095-4.

[31] M. Patrick E, N. Julius. Kruskal-Wallis Test. Wiley Online Library. pp. 1–1. https://doi.org/10.1002/9780470479216.corpsy0491. Accessed on: January 18, 2023.

[32] A. Hervé, W. Lynne, J. Newman-Keuls test and Tukey test, Encycl. Res. Des. v. 2 (2010) 897–902.