

# A Hybrid Approach for Target Discrimination in Remote Sensing: Combining YOLO and CNN-Based Classifiers

Jamesson Lira Silva<sup>1,\*</sup> , Fabiano da Cruz Nogueira<sup>1</sup> , Douglas Damião de Carvalho Honório<sup>1</sup> , Elcio Hideiti Shiguemori<sup>1</sup> , Angelo Passaro<sup>1</sup> 

<sup>1</sup>.Departamento de Ciência e Tecnologia Aeroespacial  – Instituto de Estudos Avançados – São José dos Campos (SP) – Brazil.

\*Correspondence author: jamessonjls@fab.mil.br

## ABSTRACT

With the increase in image production in recent years, there has been significant progress in the application of deep learning algorithms across various domains. Convolutional neural networks (CNNs) have been increasingly employed in remote sensing, covering all stages of target discrimination according to Johnson's criteria (detection, recognition, and identification). These CNNs are applied in many conditions and imagery from many types of sensors. In this study, we explored the use of the YOLO-v8 method, the latest version of the You Only Look Once (YOLO) family of object detection models, in conjunction with CNN architectures and supervised learning algorithms. This approach was applied to detect, recognize, and identify targets in videos captured by optical sensors, considering varying resolutions and conditions. Additionally, our research investigated the use of two CNN architectures, Inception-v3 and VGG-16, to extract relevant information from the images. The attributes obtained from the CNNs were used as input for three classification algorithms: multilayer perceptron (MLP), logistic regression, and support vector machine (SVM), thereby completing the target discrimination process. It is worth noting that in the combination of Inception-v3 and MLP, we achieved an average accuracy of 90.67%, thus completing the target discrimination process.

**Keywords:** Object recognition; Remote sensing; Neural networks; Machine learning.

## INTRODUCTION

The growing advancement of airborne and orbital sensor technologies has expanded the availability of images and videos, many of which are freely accessible (INPE 2023; USGS 2023). Data collected by satellites can reach terabytes per day, generating nearly 5 petabytes of images in 2019 (Gamba *et al.* 2011; Soille *et al.* 2018). Nanosatellites and CubeSats have become increasingly important for Earth observation, expanding the possibilities of information collection (Nagel *et al.* 2020).

**Received:** Oct. 19, 2023 | **Accepted:** Sep. 30, 2024

**Section editor:** Luiz Martins-Filho 

**Peer Review History:** Single Blind Peer Review.



However, the abundance of data poses challenges of storage, processing, and analysis, demanding efficient solutions (Gomes *et al.* 2020). To circumvent the limitation of powerful computational resources, such as general-purpose graphical processor units (GPGPUs), the use of transfer learning in the context of deep learning has become increasingly common in research across various fields (Wang *et al.* 2019). This approach allows leveraging knowledge acquired by pre-trained models on related tasks and applying them to new problems, saving time and resources.

In one such study, Rashid *et al.* (2020) employed four sets of publicly available images, including Caltech-1001 and Cifar-100. They applied two deep learning techniques to extract useful features from the images and then combined these features using data fusion methods. Recent works have been using the concept of transfer learning in areas where data is scarce, such as the medical field using X-ray, electroencephalogram, and magnetic resonance imaging (Saber *et al.* 2021; Valverde *et al.* 2021; Wan *et al.* 2021), with the aim of efficiently aiding in the detection and automatic diagnosis of various illnesses based on different examination results.

Regarding target detection and identification, Bo *et al.* (2021) conducted a comparative study, comparing statistical techniques for territorial exclusion based on the possible location of a vessel with the application of different computer vision techniques. These techniques included training a convolutional neural network (CNN) and implementing the You Only Look Once (YOLO) method. To facilitate autonomous vessel navigation, Kim *et al.* (2018) proposed the use of a deep learning-based object detection model (Faster R-CNN) to identify and locate nearby vessels in a specific region. This approach aimed to enhance the vessel's ability to operate autonomously by detecting and recognizing other ships in its vicinity.

In the task of aircraft detection (Xu and Wu 2020), using visible spectrum images with ground objects, including many types of aircraft, they proposed to enhance the YOLO method (Redmon and Farhadi 2018) using the DenseNet network (Huang *et al.* 2017). This improvement allowed for both the method's structure enhancement and improved detection accuracy. However, due to the complexity of the DenseNet structure and the large number of parameters to be extracted, this approach significantly increased the computational load of the proposed model.

For ground aircraft detection, Kharchenko and Chyrka (2018) applied the YOLO-v3 method and a simplified version called YOLO-v3 Tiny. This approach utilized videos obtained from unmanned aerial vehicles. Yang *et al.* (2021) also used YOLO-v3 for ground aircraft detection using the remote sensing object detection (RSOD) dataset (Long *et al.* 2017), consisting of publicly available images. In their experiment, they introduced an intersection over union (IoU) ratio between the region of interest in an image identified for a specific purpose, defined as real and predicted regions. The mentioned articles presented promising results; however, they used images of parked aircraft for the experiment and focused only on the first stage of the target discrimination cycle presented by John Johnson's criteria (Sjaardema *et al.* 2015), simplifying the work of the chosen network.

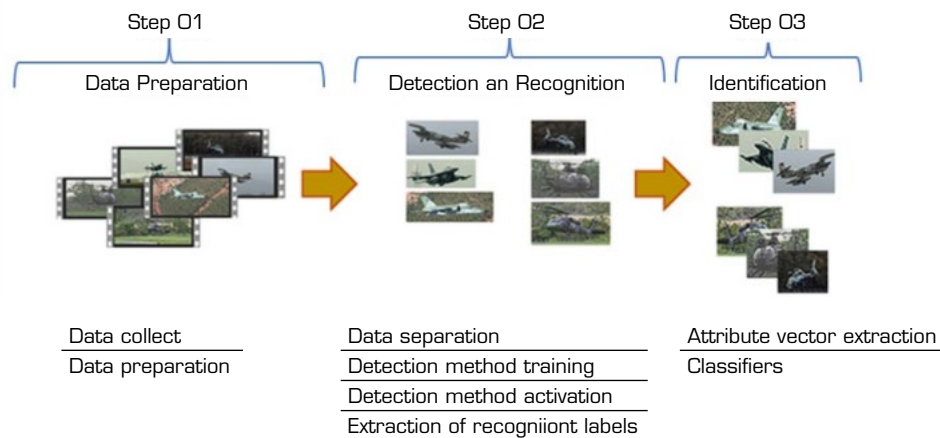
Despite the diverse applications of CNNs, the growing use of transfer learning techniques, and the promising results in object detection using different versions of the YOLO method, the combined use of these tools has not yet been explored. The model proposed in this research aims to integrate these computer vision techniques, providing an alternative that allows classifying an ever-increasing amount of data with higher accuracy, requiring less computational effort and processing time.

Considering the existence of a previously established database, the inclusion of images from content-sharing platforms was necessary to enrich the dataset. The specific choice of helicopters and airplanes as the equipment of interest was guided by the goal of the work, focused on the defense sector. In the first and second stages of the target discrimination cycle, the eighth version of the YOLO method (Terven *et al.* 2023; Ultralytics 2023) was used to detect and distinguish targets into two categories. Subsequently, to individualize the targets within the categories identified in the previous stage, attribute vectors were obtained, which are numerical representations capable of capturing image characteristics (Pan and Yang 2010). These vectors were extracted from the frames generated during the detection process using the YOLO-v8 method. For result comparison, two CNNs were used for attribute extraction: Inception-v3 (Szegedy *et al.* 2016) and VGG-16 (Simonyan and Zisserman 2014). The choice of these CNNs was grounded in their ability to extract complex image features, with a focus on the computational efficiency of Inception-v3 and the simplicity observed in the implementation of VGG-16 (Simonyan and Zisserman 2014;

Szegedy *et al.* 2016). In the final stage of the process, target identification was conducted. The attribute vectors extracted by CNNs were subjected to three different machine learning algorithms, also implemented in Orange: multilayer perceptron (MLP) (Rosenblatt 1958), support vector machine (SVM) (Boser *et al.* 1992), and logistic regression (Berkson 1944). The choice of these classifiers was based on the results achieved by Falqueto *et al.* (2019) plus the most commonly used machine learning algorithm currently (Driss *et al.* 2017). Subsequently, the results from combinations of the CNNs used in attribute extraction and the machine learning algorithms were analyzed, considering the average accuracies obtained after 50 repetitions.

## METHODOLOGY

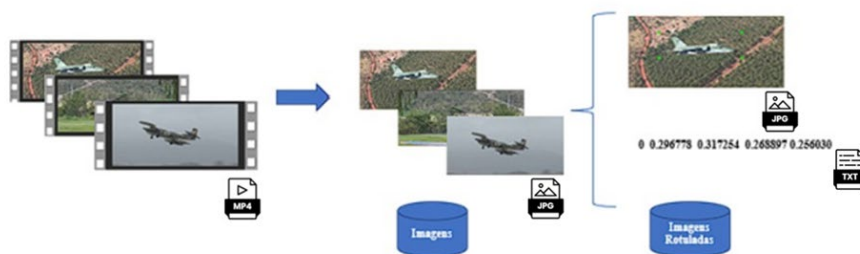
In this section, the methodology used in this work is presented, detailing the procedures adopted for the development of the experiments. The methodology consists of well-defined steps, from data collection and preparation to target identification, as can be seen in Fig. 1.



**Figure 1.** Methodology followed in this work.

The first stage of the methodology concerned over images extracted from videos in the visible spectrum, which were obtained from video-sharing platforms. These videos have different resolutions and characteristics, ranging from color to black and white. From these videos, frame extraction procedures were performed, resulting in a primary dataset.

Following the completion of the data collection process, the data preparation procedure was initiated, aiming for the detection and recognition stage using the detection method. This stage involved labeling the images in the primary database, resulting in the formation of a dataset of labeled images, as shown in Fig. 2.



**Figure 2.** Generation of labeled image databases.

The image labeling process involves assigning labels to them, which allows associating the targets present in the images with their respective locations. In the experiments chapter, a more detailed explanation of the formation of these data sets will be presented.

After the completion of the image labeling process, the resulting dataset undergoes the detection and recognition stage. The objective of this stage was to determine the location of the targets present in the scene and categorize them into specific categories. As a result, two additional sets of images were generated, representing airplanes and helicopters, as shown in Fig. 3.



Source: Elaborated by the authors.

**Figure 3.** Generation of labeled image databases.

Therefore, the following datasets were used for the completion of this work, as described in Fig. 3:

Primary dataset is the dataset obtained from videos and composed of images separated into two categories: one representing airplanes, and the other, helicopters. It is important to note that the images of these two categories are mixed within this dataset.

Labeled images is the set composed of the primary dataset plus text files containing the coordinates of object locations in their respective frames.

Airplane set is composed of recognition labels extracted from the images in the primary dataset after the detection and recognition step. It consists of six classes of airplanes and will be used for attribute vector extraction.

Helicopter set is composed of recognition labels extracted from the images in the primary dataset after the detection and recognition step and consists of six classes of helicopters to be used for attribute vector extraction.

The datasets generated by extracting recognition labels are essential during the identification step, where they are used as input for the chosen CNNs. The objective is to extract feature vectors contained in the images using the transfer learning technique. After completing this process, the extracted attributes will be used in the identification step, where previously selected machine learning algorithms are applied. These algorithms are responsible for analyzing the attributes and identifying patterns or relevant characteristics, thus performing the target identification and concluding the target discrimination cycle.

## EXPERIMENTS, RESULTS AND DISCUSSIONS

The experiments cover from data collection to evaluation of results obtained through different combinations of CNNs (for feature vector extraction) and classifier algorithms.

### Database preparation

For image collection, several videos containing the classes to be detected and recognized were examined, and frames were arbitrarily selected that exhibited at least one of the categories, allowing the inclusion of multiple targets of the same class in a single frame.

In the case of Fig. 4, we have a low-resolution video with only  $426 \times 240$  pixels.



Source: Elaborated by the authors.

**Figure 4.** Cropping for video with a resolution of  $426 \times 240$  pixels.

On the other hand, in Fig. 5, we have another example of a video used in this research, characterized by higher visual quality, with a resolution of  $3,840 \times 2,160$  pixels.



Source: Elaborated by the authors.

**Figure 5.** Cropping for video with a resolution of  $3,840 \times 2,160$  pixels.

Furthermore, in Fig. 6, we have another example of a video used, this time in adverse conditions, such as at dusk.



Source: Elaborated by the authors.

**Figure 6.** Cropping for video with adverse lighting conditions.

The analysis of the presented examples reveals the use of images with various resolutions and lighting conditions, demonstrating the methodology's ability to deal with scenarios involving variations in angles, light, and contrast. This capability is essential to ensure the effectiveness of the image detection and recognition process in different environments and situations.

During the frame extraction process, the VideoLAN (2001) media player was employed. This choice was based on its ability to efficiently extract frames from videos while maintaining the original quality. During the frame collection procedure, the images were renamed with the corresponding class name, aiming to facilitate subsequent sorting into their respective categories after the recognition label extraction.

During the frame extraction process, the primary dataset was created, consisting of two distinct categories: airplanes (Fig. 7) and helicopters (Fig. 8). Each category comprised 3,000 images, resulting in a dataset containing 6,000 images.



Source: Elaborated by the authors.

**Figure 7.** Aircraft images in the primary dataset.



Source: Elaborated by the authors.

**Figure 8.** Helicopter images in the primary dataset.

As illustrated in Fig. 9, the database is divided into two categories: Airplane and Helicopter. Each category consists of different aircraft classes. In the Airplane category, we have six distinct classes: P-3 AM Orion, KC-390, AMX, AT-37, Boeing 777, and F-16. Meanwhile, the Helicopter category is subdivided into six classes: Écureuil, Black Hawk, BO-105, Chinook, Panther, and Super Puma.



Source: Elaborated by the authors.

**Figure 9.** Target classes.

In the next step, the process of labeling the images in the primary dataset was initiated with the purpose of training YOLO. To ensure that the coordinates were not altered, it was decided to resize the images to a resolution of  $640 \times 640$  pixels (Bochkovskiy *et al.* 2004; Wang *et al.* 2021) before starting this step.

### Labeling process

To carry out the image labeling process, the LabelImg framework (Tzutalin 2015) was used, which is one of the most employed tools for image labeling tasks in YOLO applications (Andrade 2022; Dharneeshkar *et al.* 2020; Muksit *et al.* 2022; Varnima and Ramachandran 2020). This choice was made because it is a graphic annotation tool for images that allows offline labeling, is free and open-source, available on GitHub, and enables saving work in various formats, including YOLO.

After the labeling process was completed, a dataset of labeled images was generated, exemplified in Fig. 10, containing the images from the primary set, accompanied by text files that record the coordinates of the areas of interest in the respective images, as well as an additional file containing the present categories.



Source: Elaborated by the authors.

**Figure 10.** Labeled images.

### Detection and recognition of targets

With the aim of improving processing and reducing the computational load during the YOLO method training, a strategy was adopted involving the division of the labeled image dataset into three subsets, following the proportion of 50/40/10, as presented in Table 1. These subsets were defined as the training set, validation set, and test set, comprising 3,000, 2,400, and 600 labeled images, respectively.

**Table 1.** Labeled image dataset distribution.

Element	Subsets	Number of images per subsets
Set of labeled images	Training	3,000
	Validation	2,400
	Testing	600

Source: Elaborated by the authors.

In addition, representative videos from both studied categories were used as supplementary examples in the activation process.

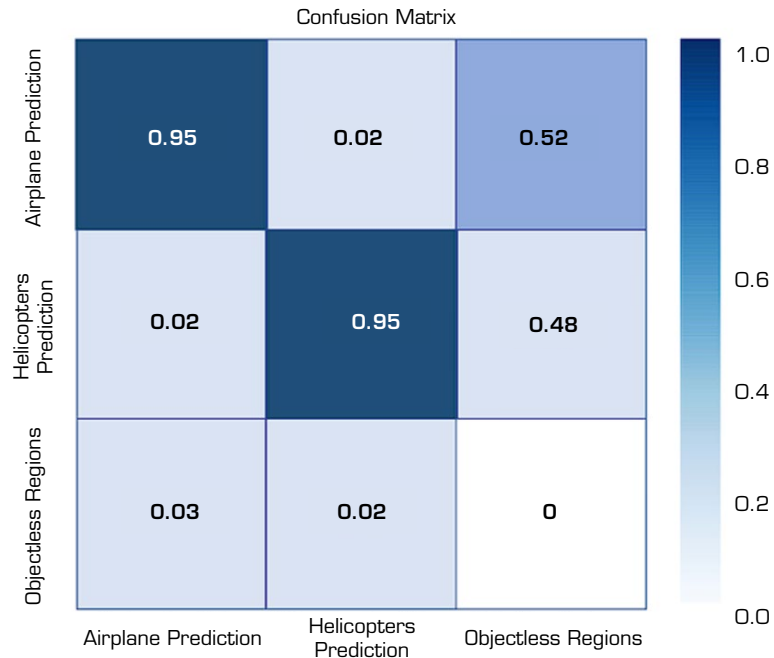
### *Results of the detection and recognition stage*

The YOLO method was trained and validated using a dataset composed of 10,801 files, containing both images and labels with target coordinates, over 24 epochs as conventionally chosen in this study.

The YOLO training process was carried out using advanced computational resources, including the utilization of parallel computing architectures and GPUs, within the collaborative environment of Google Colab. The training process required a total time of 8 hours and 7 minutes to be completed. This duration was considered reasonable for obtaining satisfactory results, considering the dataset characteristics and the available computational capacity.



In Fig. 11, the confusion matrix resulting from the validation process of the method is presented. The columns represent the predictions made by the network, while the rows represent the actual objects. The analysis of the confusion matrix reveals promising results in the detection and recognition phase, with an average accuracy of 95% in the airplane dataset and 96% in the helicopter dataset.



Source: Elaborated by the authors.

**Figure 11.** Confusion matrix obtained on the YOLO test set.

Figure 12 shows examples of images obtained during the detection and recognition step.



Source: Elaborated by the authors.

**Figure 12.** Detection and recognition stage.

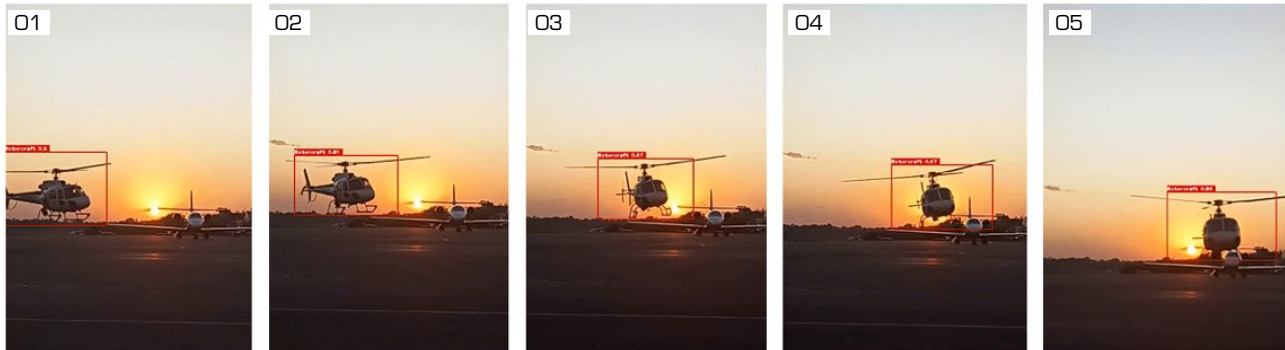
During video activation, the neural network exhibited consistent behavior in the process of target detection and recognition. Even in the face of variations in viewing angles (Fig. 13), adverse lighting scenarios (Fig. 14), or situations involving multiple targets and cases of occlusion (Fig. 15), the network successfully detected and recognized the trained targets.





Source: Elaborated by the authors.

**Figure 13.** Videos after detection and recognition step – different angles.



Source: Elaborated by the authors.

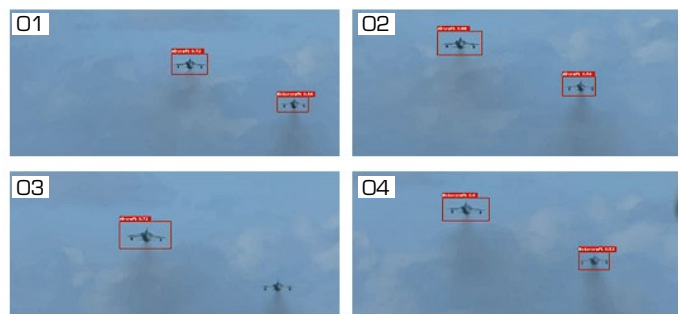
**Figure 14.** Videos after the detection and recognition step – adverse scenario.



Source: Elaborated by the authors.

**Figure 15.** Videos after the detection and recognition step – adverse scenario.

However, it was observed that in some circumstances, there were failures in detection or errors in recognition, as evidenced in Fig. 16, where moments of failure were identified, including cases where objects were not detected or where detection was correct, but recognition failed.



Source: Elaborated by the authors.

**Figure 16.** Videos with failures in the detection and recognition process.



### *Extraction of recognition labels*

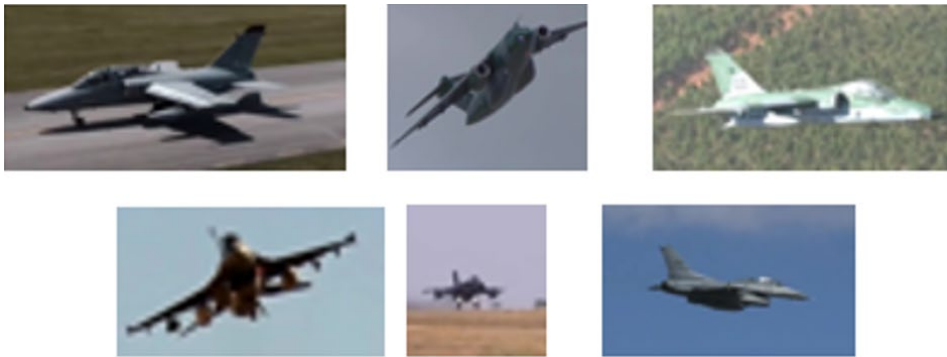
During YOLO activation, the recognition label extraction process was also carried out (Fig. 17). This process aimed to extract relevant target-related information, thereby eliminating extraneous information. The purpose of this extraction is to prepare the data for the subsequent target identification step.



Source: Elaborated by the authors.

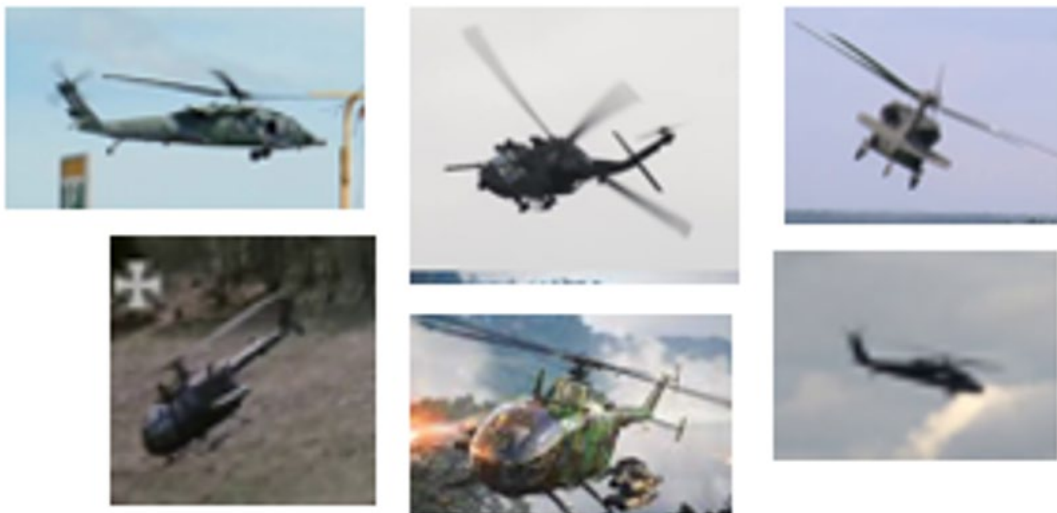
**Figure 17.** Extraction of recognition labels.

During this process, a set consisting of 5,783 images representing the regions of interest obtained during the activation of the detection method was created. These images were divided into two distinct categories: one category containing images of airplanes (Fig. 18) and another containing images of helicopters (Fig. 19).



Source: Elaborated by the authors.

**Figure 18.** Aircraft recognition labels.



Source: Elaborated by the authors.

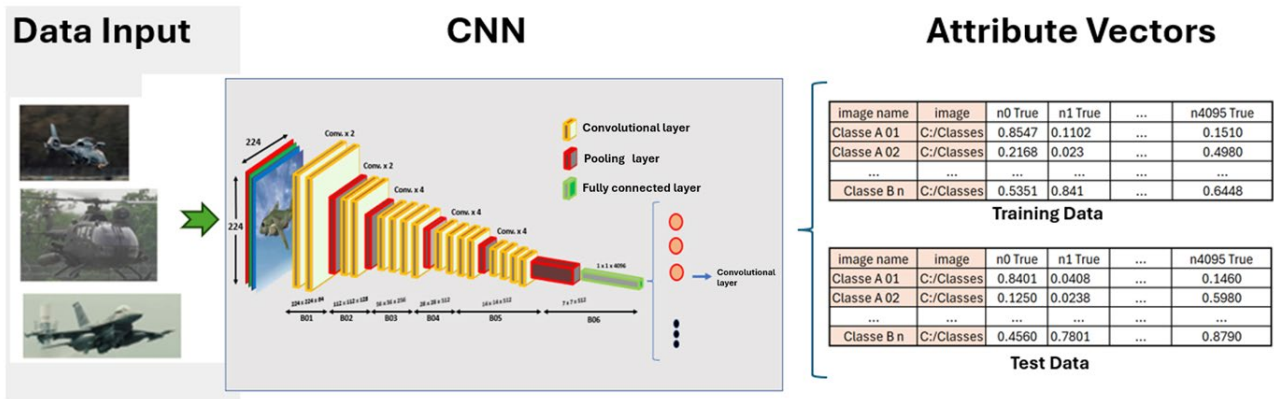
**Figure 19.** Helicopter recognition labels.

The generated datasets will be used as the database to feed the CNNs, aiming to extract the feature vectors employed in the target identification stage.

### Target identification

To extract the attribute vectors, the airplane dataset and the helicopter dataset, each composed of 5,783 images, were used. These datasets were generated through the label extraction process for recognition.

For the extraction of attribute vectors (Fig. 20), the pre-implemented CNNs Inception-v3 and VGG-16 were utilized within the Python package Orange (Orange Data Mining).



Source: Elaborated by the authors.

**Figure 20.** Feature extraction process.

These employed CNNs were pretrained on the ImageNet database, enabling a quick classification process with reduced computational effort and a smaller number of images.

In order to obtain a balanced dataset for identification and achieve statistically more reliable results, an adaptation of the cross-validation technique was applied. Table 2 presents the random subdivision performed, in which 1,500 attribute vectors were separated per category, with 250 vectors per class, from an initial set of 5,783 attribute vectors. This division was planned to allocate 80% of the data for the algorithm’s training phase, while the remaining 20% was reserved for testing to evaluate the method’s accuracy.

**Table 2.** Data distribution for classification.

Repetitions	Element	Structure sequential	Number of images per class	Total
50	Airplane	Training	200	1,200
		Testing	50	300
		Total		1,500
	Helicopters	Training	200	1,200
		Testing	50	300
		Total		1,500

Source: Elaborated by the authors.

After completing the attribute vector extraction process, the following numbers of attributes were obtained for each CNN, as shown in Table 3.



**Table 3.** Information extracted by CNNs.

CNN	Extracted information	Quantity
Inception-v3	Attributes	2,048
VGG-16	Attributes	4,096

Source: Elaborated by the authors.

For the identification step, the following classifiers were used: MLP, logistic regression, and SVM. Since the Orange library was used, there was no concern about optimizing the algorithm specifications used.

### Identification stage

After performing the 50 repetitions, the results of the classifications of the attribute vectors extracted by the Inception-v3 and VGG-16 networks, using the neural network, SVM, and logistic regression classifiers were computed and presented in Tables 4 and 5.

**Table 4.** Average accuracy using the airplane dataset.

Class A – Airplane (accuracy)			
CNN	MLP	Logistic regression	SVM
Inception-v3	0.9067	0.8947	0.8777
VGG-16	0.7758	0.7911	0.7693

Source: Elaborated by the authors.

**Table 5.** Average accuracy using the helicopter dataset.

Class A – Helicopter (accuracy)			
CNN	MLP	Logistic regression	SVM
Inception-v3	0.8405	0.8299	0.8210
VGG-16	0.7691	0.7695	0.6984

Source: Elaborated by the authors.

Upon analyzing the presented results, it is possible to obtain relevant preliminary information. It should be noted that the application of the Inception-v3 CNN as a feature extractor resulted in the best performance achieved. Additionally, it was found that the dataset consisting of airplane images outperformed the helicopter dataset, regardless of the employed CNN. Furthermore, it was observed that the two most promising classifiers were MLP and logistic regression. These algorithms showed significant results when using both the airplane and helicopter datasets.

Since the data used in this stage are balanced and there is no emphasis on any specific class, accuracy was used as a metric for generating graphs to demonstrate the results obtained in the above-mentioned combinations. These graphs provide a visualization of the accuracy distribution over the 50 repetitions conducted, allowing for a more comprehensive and detailed analysis of the results obtained.

In order to highlight the most promising combinations, a more in-depth analysis of the results was conducted. In this analysis, attribute vectors extracted by Inception-v3 were used in conjunction with MLP and logistic regression classifiers, generating confusion matrices from a single random repetition. The purpose of this approach was to examine the behavior of the algorithms that showed the best performance, with a view to their practical application.

When analyzing the achieved results (Tables 4 and 5), it becomes evident that combinations using Inception-v3 yielded superior results compared to when VGG-16 was used as the feature extractor, both in the aircraft and helicopter datasets.

Analyzing one of the repetitions performed using MLP to classify attributes extracted by Inception-v3 (Fig. 21), it was observed that the B-777 class exhibited a high accuracy rate, reaching 0.98. On the other hand, the F-16 class showed the

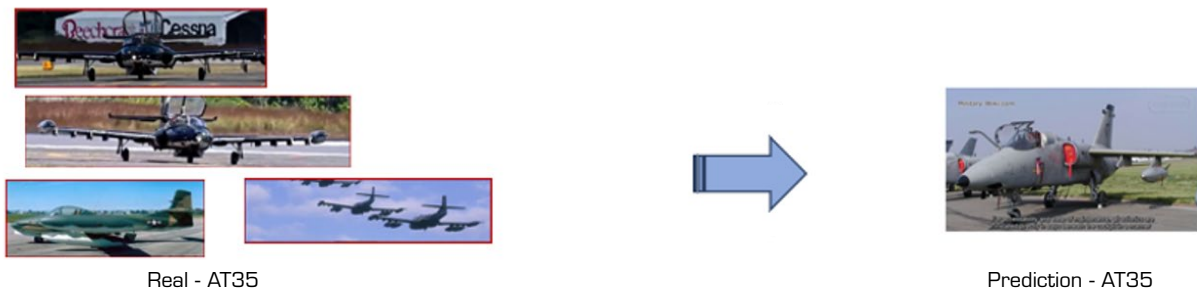
lowest performance, with an accuracy of 0.86. These results indicate a noticeable similarity in the accuracies obtained for the six analyzed classes.

		Prediction						Σ
		AMX	AT 35	B777	F16	KC390	P3	
Truth	AMX	49	0	0	1	0	0	50
	AT 35	4	46	0	0	0	0	50
	B777	0	1	48	0	0	1	50
	F16	2	3	0	43	0	2	50
	KC390	1	1	0	1	47	0	50
	P3	1	0	3	2	0	44	50
Σ		57	51	51	47	47	47	300

Source: Elaborated by the authors.

**Figure 21.** Confusion matrix in one of the repetitions with the MLP (aircraft).

Continuing the analysis of the confusion matrix related to the classification of the airplane image dataset, the AT-35 was misidentified on four occasions, all classified as AMX. These mistaken results are illustrated in Fig. 22, where two images are frontal, and one of them contains more than one aircraft in the scene, although all airplanes are AT-35.



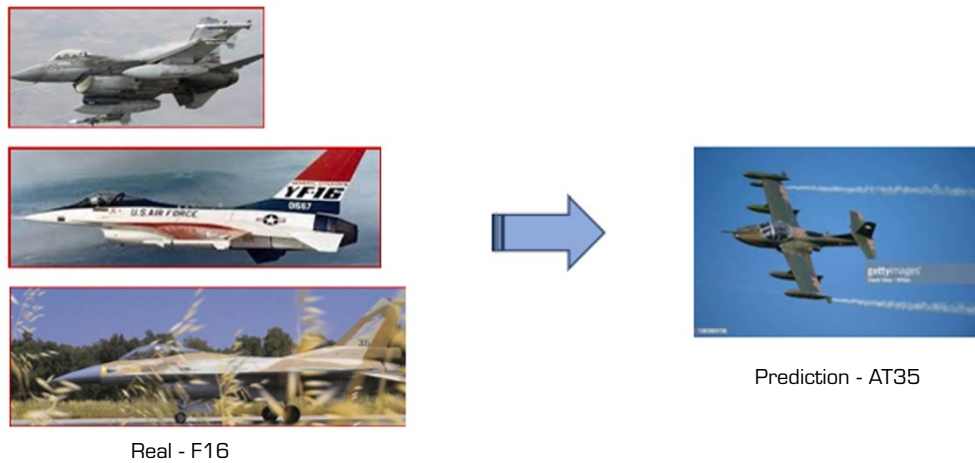
Source: Elaborated by the authors.

**Figure 22.** Images of the AT-35 aircraft with identification errors.

In cases where there are multiple airplanes, identification indeed becomes a challenging task, as it requires dealing with the presence of multiple objects in the scene. However, the first scenario may indicate a scarcity of images of the mentioned aircraft in the specified position in the training dataset.

In Fig. 23, three images with an identification error between F-16 and AT-35 were presented. It is possible to observe that these airplanes do not bear significant resemblance to each other, with no evident justification in the images themselves for this error. Thus, it becomes necessary to adopt some strategies, such as adjusting hyperparameters, to refine the network or increasing the training dataset so that the model can capture the distinct characteristics of the mentioned airplanes.





Source: Elaborated by the authors.

**Figure 23.** Images of the F-16 airplane with identification errors.

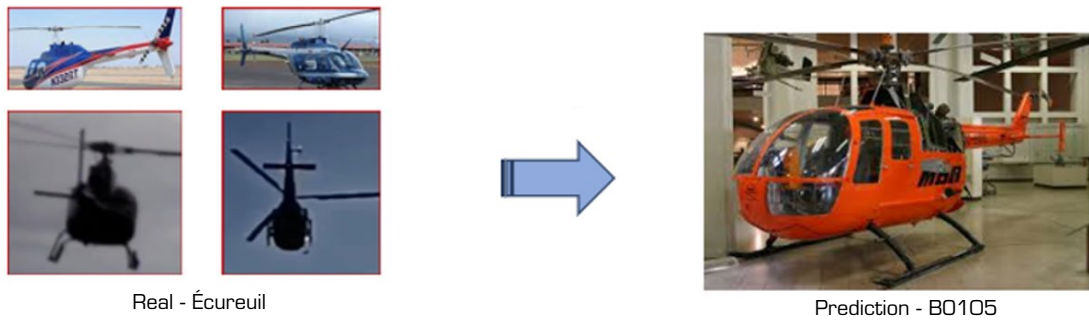
In the analysis of the confusion matrix, which portrays the results of applying the MLP to the set of attributes extracted from the helicopter dataset (Fig. 24), a high confusion rate was observed in the identification between the Écureuil and BO-105 helicopters, as well as between the Super Puma and Panther.

Truth	Prediction							$\Sigma$
	B105	Black Hawk	Chinook	Écureuil	Panther	Super Puma		
B105	43	2	1	3	1	1	50	
Black Hawk	0	42	1	0	4	3	50	
Chinook	0	0	44	1	2	2	50	
Écureuil	7	0	1	34	5	3	50	
Panther	0	1	0	0	47	1	50	
Super Puma	1	0	0	1	8	39	50	
$\Sigma$	52	46	47	39	67	49	300	

Source: Elaborated by the authors.

**Figure 24.** Confusion matrix in one of the repetitions with the MLP (helicopters).

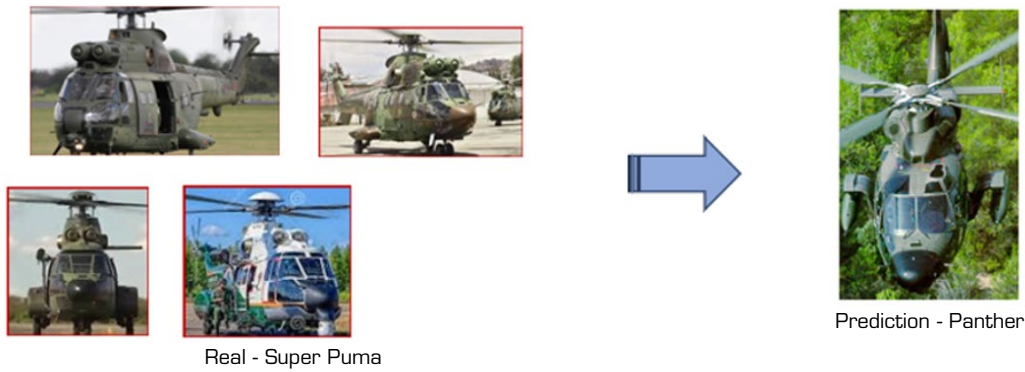
Through Fig. 25, we can observe that the images showing errors in the identification of the Écureuil helicopter are characterized by low definition, making precise identification of the aircraft challenging. Additionally, it is important to note that the Écureuil helicopter shares some similarities with the BO-105 helicopter, which can further increase the difficulty in correctly distinguishing these two aircraft.



Source: Elaborated by the authors.

**Figure 25.** Images of the Écureuil helicopter with identification errors.

In relation to identification errors between Super Puma and Panther helicopters, it can be observed in Fig. 26 that the majority of errors are related to frontal images of the helicopters, at which point they are most similar, indicating a need to enrich the training database with this specific view.



Source: Elaborated by the authors.

**Figure 26.** Images of the Super Puma helicopter with identification errors.

The results obtained through the application of logistic regression and MLP showed significant similarities in the values during the identification process, as observed in Tables 4 and 5. However, a more detailed analysis reveals differences between the behaviors of the two algorithms, especially in the analysis of the confusion matrix of one of the repetitions related to the aircraft dataset (Fig. 27).

Truth	Prediction							Σ
	AMX	AT 35	B777	F16	KC390	P3		
AMX	49	0	0	1	0	0		50
AT 35	3	45	0	0	0	2		50
B777	0	1	48	0	0	1		50
F16	5	2	0	41	0	2		50
KC390	1	1	1	1	46	0		50
P3	1	0	3	1	0	45		50
Σ	57	51	51	47	47	47		300

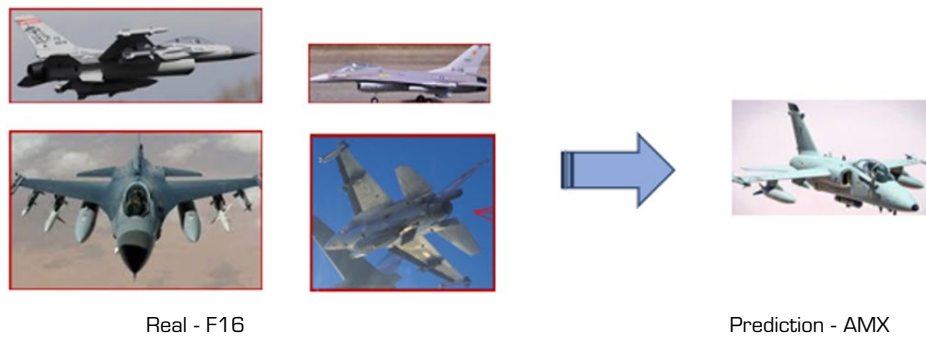
Source: Elaborated by the authors.

**Figure 27.** Confusion matrix in one repetition using logistic regression (aircraft).



While the MLP had difficulties in identifying the AT-35 aircraft, incorrectly classifying it as an AMX (Fig. 24), logistic regression faced greater challenges in classifying the F-16 (Fig. 27). These discrepancies reveal distinct patterns in the behavior of the classifiers, even when their performances are very close to each other.

Regarding errors in the classification of the F-16, it can be observed in Fig. 28 that the images in which the network incorrectly classified the F-16 as an AMX presented challenging characteristics, such as visual similarities and different perspectives, making it difficult to correctly identify the target.



Source: Elaborated by the authors.

**Figure 28.** Images of the F-16 aircraft with identification errors.

During the application of the logistic regression method for helicopter identification, an additional difficulty was observed in distinguishing between the Écureuil and BO-105 models, as described in the confusion matrix presented in Fig. 29.

Truth	Prediction							$\Sigma$
	B105	Black Hawk	Chinook	Écureuil	Panther	Super Puma		
B105	42	3	0	3	1	1	50	
Black Hawk	0	41	1	1	4	3	50	
Chinook	0	1	44	1	2	2	50	
Écureuil	7	0	0	36	5	2	50	
Panther	2	1	0	1	44	2	50	
Super Puma	5	0	0	0	8	39	50	
$\Sigma$	52	46	47	39	67	49	300	

Source: Elaborated by the authors.

**Figure 29.** Confusion matrix in one repetition using logistic regression (helicopters).

In Fig. 30, images are presented that reveal flaws in the identification performed by the logistic regression classifier when attempting to identify the Écureuil helicopter.

The results presented in Table 6 offer a comparison between the performance of MLP and logistic regression in the aircraft classification task. The MLP demonstrated slightly superior performance in most classes, particularly in the identification of the F-16, with a precision of 0.86 compared to 0.82 for logistic regression. However, logistic regression achieved better precision in the P3 class (0.90 vs. 0.88).





Source: Elaborated by the authors.

**Figure 30.** Images of the Écureuil helicopter with identification errors.

**Table 6.** Comparison of accuracy between MLP and logistic regression in the aircraft discrimination task.

Class	Precision MLP	Precision regression logistic
AMX	49/50 = 0.98	49/50 = 0.98
AT35	46/50 = 0.92	45/50 = 0.90
B777	48/50 = 0.96	48/50 = 0.96
F-16	43/50 = 0.86	41/50 = 0.82
KC390	47/50 = 0.94	46/50 = 0.92
P3	44/50 = 0.88	45/50 = 0.90

Source: Elaborated by the authors.

In the helicopter classification task, as shown in Table 7, there were notable differences between the two models, particularly in the Écureuil class, where logistic regression achieved better precision (0.72 vs. 0.68). However, the MLP performed better in the Panther class (0.94 vs. 0.88). Both models performed equally well in the Chinook and Super Puma classes.

**Table 7.** Comparison of accuracy between MLP and logistic regression in the helicopter discrimination task.

Class	Precision MLP	Precision regression logistic
B105	43/50 = 0.86	42/50 = 0.84
BlackHawk	42/50 = 0.84	41/50 = 0.82
Chinook	44/50 = 0.88	44/50 = 0.88
Écureuil	34/50 = 0.68	36/50 = 0.72
Panther	47/50 = 0.94	44/50 = 0.88
Super Puma	39/50 = 0.78	39/50 = 0.78

Source: Elaborated by the authors.

It is possible to observe that, despite the helicopters being similar, the network failed to capture the details to the extent of performing a reliable identification of the Écureuil helicopter, leading to errors that could have been avoided by adjusting hyperparameters or increasing the database.

### Discussions and future work

From the obtained results, it is possible to make some observations that may encourage future research and improvements to the employed models. In the research in question, for the data used, the YOLO method demonstrated promising performance in object detection and recognition, achieving an average accuracy of 95% in the airplanes category and 96% in the helicopters category. It has been observed that the methodology achieved promising practical results in the task of



detection and recognition even in adverse lighting conditions, different viewing angles, or when applied to images of varying resolutions. The method's demonstrated ability to tackle challenges in complex scenarios reinforces its effectiveness in target detection and recognition applications.

Despite these positive results, some failures in the detection and recognition stage were also observed. These occurrences represent issues that need to be addressed to ensure the optimal performance of the network in different circumstances. Detection failures may have been caused by low image resolution, inadequate lighting, or challenging viewing angles, potentially leading to network confusion. Recognition errors, on the other hand, can arise due to visual similarities between different objects, variations in the appearance of these objects, or even the lack of sufficient training data to correctly categorize certain object classes.

To overcome these limitations, it is necessary to adopt some strategies to enhance the research. For example, a broader and more diversified data collection can be carried out, aiming to more accurately represent the various situations in which the network will be applied. Furthermore, it is necessary to conduct a detailed analysis of the cases in which failures occurred, aiming to include possible examples of images with these issues in the network's training, in order to fill this knowledge gap. Finally, the use of an infrastructure that allows for an increase in training time, with the aim of increasing the number of epochs, can also be adopted as part of the strategies to improve this research phase.

Regarding the applied CNNs, the significant impact of the Inception architecture on the task of extracting image attributes stands out when compared to the VGG network family. Although it extracts a lower number of attributes (2,048) compared to VGG-16, Inception-v3 excelled in the task of extracting discriminative and relevant attributes from the images. The technique of parallel convolutions with different filter sizes employed by Inception-v3 may have assisted in capturing information from different viewing angles and images of different resolutions, allowing the network to learn richer and more robust representations of objects in the images.

Another factor that may be related to the results obtained in the classification of the extracted attributes is the depth of the networks used as feature extractors. While Inception-v3 has 48 layers, VGG-16 has 16. The depth is related to the CNN's ability to learn complex and hierarchical representations of data. The greater depth of Inception-v3 may have allowed the capture of more significant features, which can be advantageous for the task of image classification.

Regarding the classifier algorithms, one relevant observation to highlight in the results obtained is the good performance achieved by MLP, logistic regression, and SVM compared to the other classifiers. One of the reasons for the better results obtained by these algorithms, such as MLP and logistic regression, may be attributed to their generalization capability. Both neural networks and logistic regression are known for their good generalization capability.

After analyzing the confusion matrices generated by the classifiers on the aircraft database (Figs. 21 and 27) and helicopters (Figs. 24 and 29), it was observed that the model demonstrated consistency in terms of sensitivity, specificity, and precision in most classes. This suggests that the model correctly identifies the majority of instances in the experiment, with low rates of false positives. However, exceptions were noted in the F-16, P3, Écureuil, and Super Puma classes, which exhibited a relatively higher proportion of false positives compared to other classes. These results indicate the possibility of adjusting the parameters of the classifiers as a measure to enhance the overall performance of the model.

It is necessary to keep in mind that the selection of the most suitable classifier algorithm depends on the specific problem, the available dataset, the characteristics of the variables, and other relevant factors. It is important to emphasize the importance of conducting experiments and tests with various algorithms to determine which one is most suitable for the specific problem at hand. Furthermore, each algorithm has its advantages and disadvantages, and it is recommended to evaluate and compare their performance in relation to specific data and objectives.

Additionally, it was observed that in all combinations of CNNs and classifiers, it is possible to note that the evaluation metrics achieved higher values in experiments using attributes extracted from the category of airplanes. This suggests that the classification of the category of helicopters is more challenging, which can be explained by the specific geometric characteristics of these devices.

For future research, we suggest exploring the application of the proposed methodology with different combinations of CNNs and machine learning algorithms. Furthermore, it is crucial to consider the full implementation of YOLO-v8, covering the entire target discrimination process. It is also relevant to evaluate the performance of other real-time detection networks, such as EfficientDet, to understand the actual benefits of the approach proposed in this study.

## CONFLICT OF INTEREST

Nothing to declare.

## AUTHORS' CONTRIBUTION


**Conceptualization:** Silva JL, Shiguemori EH, Passaro A; **Methodology:** Silva JL, Shiguemori EH, Passaro A; **Resources:** Silva JL; **Investigation:** Silva JL; **Data curation:** Silva JL, Nogueira FC; **Supervision:** Passaro A; **Original – draft writing:** Silva JL; **Writing – review and editing:** Nogueira FC, Honório DDC, Shiguemori, EH Passaro A; **Final approval:** Passaro A.

## DATA AVAILABILITY STATEMENT

The data will be available upon request.

## FUNDING

Conselho Nacional de Desenvolvimento Científico e Tecnológico   
Grant No: 307691/2020-9

Coordenação de Aperfeiçoamento de Pessoal de Nível Superior   
Finance Code 001

## ACKNOWLEDGMENTS

To Brazilian Air Force for supporting this study.

## REFERENCES

[INPE] Instituto Nacional de Pesquisas Espaciais (2023) Catálogo de imagens. [accessed Jul 25 2023]. [dgi.inpe.br/catalogo/explore](http://dgi.inpe.br/catalogo/explore)

[USGS] United States Geological Survey (2023) Earth Explorer. [accessed Jul 25 2023]. <https://earthexplorer.usgs.gov/>

[VLC] VideoLAN Organization (2001) Paris, École Centrale Paris Université. [accessed Sep 29 2022]. <https://www.videolan.org/videolan/>

Andrade RM (2022) Detecção de comportamentos de veículos a partir de imagens de drones e de monitoramento (master's thesis). São José dos Campos: Instituto Nacional de Pesquisas Espaciais. In Portuguese. Available in: <http://mtc-m21d.sid.inpe.br/col/sid.inpe.br/mtc-m21d/2022/07.20.19.46/doc/publicacao.pdf?metadatarpository=sid.inpe.br/mtc-m21d/2022/07.20.19.46.42&mirror=urllib.net/www/2021/06.04.03.40.25&languagebutton=pt-BR>



- Berkson J (1944) Application of the logistic function to bio-assay. *J Am Stat Assoc* 39(227):357-365. <https://doi.org/10.1080/01621459.1944.10500699>
- Bo LI, Xiaoyang XIE, Tang W (2021) Ship detection and classification from optical remote sensing images: a survey. *Chinese J Aeronaut* 34(3):145-163. <https://doi.org/10.1016/j.cja.2020.09.022>
- Bochkovskiy A, Wang CY, Liao HYM (2004) Yolov4: optimal speed and accuracy of object detection. arXiv:2004:10934. <https://doi.org/10.48550/arXiv.2004.10934>
- Boser BE, Guyon IM, Vapnik VN (1992) A training algorithm for optimal margin classifiers. Paper presented 1992 Proceedings of the Fifth Annual Workshop on Computational Learning Theory; Pittsburgh Pennsylvania USA.
- Dharneeshkar J, Dhakshana V, Aniruthan SA (2020) Deep learning based detection of potholes in Indian roads using YOLO. Paper presented 2020 International Conference on Inventive Computation Technologies. IEEE; Coimbatore, India.
- Driss SB, Soua M, Kachouri R (2017) A comparison study between MLP and convolutional neural network models for character recognition. Paper presented 2017 SPIE Conference Real-Time Image and Video Processing. Society of Photographic Instrumentation Engineers; Anaheim, CA, United States.
- Falqueto LE, Sá JAS, Paes RL (2019) Oil rig recognition using convolutional neural network on Sentinel-1 SAR images. *IEEE Geosci Remote Sens Lett* 16(8):1329-1333. <https://doi.org/10.1109/LGRS.2019.2894845>
- Gamba P, Du P, Juergens C (2011) Foreword to the special issue on “human settlements: a global remote sensing challenge”. *IEEE J Sel Top Appl Earth Obs Remote Sens* 4(1):5-7. <https://doi.org/10.1109/JSTARS.2011.2106332>
- Gomes VCF, Queiroz GR, Ferreira KR (2020) An overview of platforms for big earth observation data management and analysis. *Remote Sens* 12(8):1253. <https://doi.org/10.3390/rs12081253>
- Huang G, Liu Z, van der Maaten L (2017) Densely connected convolutional networks. Paper presented 2017 Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE; Honolulu, HI, USA.
- Kharchenko V, Chyrka I (2018) Detection of airplanes on the ground using YOLO neural network. Paper presented 2018 IEEE 17th International Conference on Mathematical Methods in Electromagnetic Theory. IEEE; Kyiv, Ukraine.
- Kim K, Hong S, Choi B, Kim E (2018) Probabilistic ship detection and classification using deep learning. *Appl Sci* 8(6):936. <https://doi.org/10.3390/app8060936>
- Long Y, Gong Z, Xiao Z, Liu Q (2017) Accurate object localization in remote sensing images based on convolutional neural networks. *IEEE Trans Geosci Remote Sens* 55(5):2486-2498. <https://doi.org/10.1109/TGRS.2016.2645610>
- Muksit AA, Hasan F, Emon FHB (2022) YOLO-Fish: a robust fish detection model to detect fish in realistic underwater environment. *Ecol Inform* 72:101847. <https://doi.org/10.1016/j.ecoinf.2022.101847>
- Nagel GW, Novo EMLM, Kampel M (2020) Nanosatellites applied to optical Earth observation: a review. *Rev Ambiente Água* 15(3). <https://doi.org/10.4136/ambi-agua.2513>
- Pan SJ, Yang Q (2010) A survey on transfer learning. *IEEE Trans Knowl Data Eng* 22(10):1345-1359. <https://doi.org/10.1109/TKDE.2009.191>
- Rashid M, Khan MA, Alhaisoni M (2020) A sustainable deep learning framework for object recognition using multi-layers deep features fusion and selection. *Sustainability* 12(12):5037. <https://doi.org/10.3390/su12125037>
- Redmon J, Farhadi A (2018) Yolov3: an incremental improvement. arXiv:1804.02767. <https://doi.org/10.48550/arXiv.1804.02767>



- Rosenblatt F (1958) The perceptron: a probabilistic model for information storage and organization in the brain. *Psychol Review* 65(6):386. <https://doi.org/10.1037/h0042519>
- Saber A, Sakr M, Abo-Seida OM (2021) A novel deep-learning model for automatic detection and classification of breast cancer using the transfer-learning technique. *IEEE Access* 9:71194-71209. <https://doi.org/10.1109/ACCESS.2021.3079204>
- Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556. <https://doi.org/10.48550/arXiv.1409.1556>
- Sjaardema TA, Smith CS, Birch GC (2015) History and evolution of the Johnson criteria. Albuquerque: Sandia.
- Soille P, Burger A, Marchi D (2018) A versatile data-intensive computing platform for information retrieval from big geospatial data. *Future Gener Comput Syst* 81:30-40. <https://doi.org/10.1016/j.future.2017.11.007>
- Szegedy C, Vanhoucke V, Loffe S (2016) Rethinking the Inception architecture for computer vision. Paper presented 2016 Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE; Las Vegas, NV, USA.
- Terven J, Córdova-Esparza DM, Romero-González JA (2023) A comprehensive review of yolo architectures in computer vision: from yolov1 to yolov8 and yolo-nas. *Mach Learn Knowl Extr* 5(4):1680-1716. <https://doi.org/10.3390/make5040083>
- Tzutalin D (2015) LabelImg. GitHub repository 6(4).
- Ultralytics (2023) Introducing YOLOv8 Docs. [accessed Feb 18 2024]. <https://docs.ultralytics.com/task/classify//>
- Valverde JM, Imani V, Abdollahzadeh A (2021) Transfer learning in magnetic resonance brain imaging: a systematic review. *J Imaging* 7(4):66. <https://doi.org/10.3390/jimaging7040066>
- Varnima EK, Ramachandran C (2020) Real-time gender identification from face images using you only look once (YOLO). Paper presented 2020 4th International Conference on Trends in Electronics and Informatics. IEEE; Tirunelveli, India.
- Wan Z, Yang R, Huang M (2021) A review on transfer learning in EEG signal analysis. *Neurocomputing* 421:1-14. <https://doi.org/10.1016/j.neucom.2020.09.017>
- Wang CY, Bochkovskiy A, Liao HYM (2021) Scaled-yolov4: scaling cross stage partial network. Paper presented 2021 Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE;
- Wang Y, Wang C, Luo L (2019) Image classification based on transfer learning of convolutional neural network. Paper presented 2019 Chinese Control Conference. IEEE; Guangzhou, China.
- Xu D, Wu Y (2020) Improved YOLO-V3 with DenseNet for multi-scale remote sensing target detection. *Sensors* 20(15):4276. <https://doi.org/10.3390/s20154276>
- Yang Y, Liao Y, Cheng L (2021) Remote sensing image aircraft target detection based on GIoU-YOLO v3. Paper presented 2021 6th International Conference on Intelligent Computing and Signal Processing. IEEE; Xi'an, China.