Deep Learning for Brazilian Car Make and Model Recognition

Lucas Augusto Albini, Matheus Gutoski, Heitor Silvério Lopes
Programa de Pós Graduação em Engenharia Elétrica e Informática (CPGEI)
Universidade Tecnológica Federal do Paraná (UTFPR)
Av. 7 de setembro, 3165 - 80230-901, Curitiba (PR), Brazil
Email: {lucasalbini, matheusgutoski}@alunos.utfpr.edu.br
, hslopes@utfpr.edu.br

***Abstract*—Car make and model classification is an issue frequently discussed in the literature due to its several applications in security, traffic control, and urban planning, especially in the context of smart cities. Currently, deep learning methods are the state-of-the-art for image and video classification. This work it is presented a method for classifying cars at the level of make and model in a simple and effective way using deep learning methods. To accomplish this task, the Inception-v3 neural network was used to train and evaluate the model. Another objective of this work is to create a high-quality dataset of images of cars produced by the Brazilian industry. The full dataset has 24319 images distributed into 10 makes and 50 models, with an average of 500 images per class. The average classification accuracy reached 82.36% and 94,87%, when considering the top-3 results. Our results showed that the proposed approach was very successful for classification purposes and encourages further development.**

***Keywords*—Image Segmentation; Evolutionary Computation; Genetic Expression Programming; Color recognition;**

I. Introduction

Smart cities are those that use technology intensively to bring well-being to their population, improving security and mobility, on the one hand, favoring economic growth and, on the other, sustainability. In large Brazilian cities, there is a growing concern about mobility and security, mainly due to the growing number of traffic accidents and car thefts.

Over the years, the number of security cameras in cities has grown exponentially, whether public or private, and they are scattered all over the places. However, as a matter of fact, there are not enough people to watch them at all times. This fact leads to purely reactive systems, such that the surveillance cameras are useful only to view past events. This is a motivation for devising means of the automatic analysis of images and videos from surveillance cameras in such a way to circumvent the need for human experts.

In smart cities, cameras are used, mainly, for traffic or people surveillance. License plate recognition from images is a known technology already available as a standard product. However, identifying the car make and model is much more difficult not only due to the great variability of models but, also, due to environmental factors (sight angle, illumination, distance to the object, etc).

A system capable of recognizing the car make and model from surveillance cameras may be of great importance for public security. It could be used, for instance, to prevent thefts or increase police promptness, besides being useful for traffic control and urban planning.

Object classification in images and videos is a widely discussed issue in the literature. Many different approaches have appeared along time to tackle this problem. Despite the recent advances using Convolutional Neural Networks (CNNs), with excellent results [1], we still have problems where the objects under comparison are quite similar, for instance, in the car make and model classification [2].

This work proposes a method for classifying car make and model from raw images. In order to train the classifier, a new dataset was created, using images of vehicles recently produced in Brazil. This dataset includes images of several makes and models, with a large diversity of angles and backgrounds.

The paper was organized as follows: Section II presents theoretical aspects of the work. Section III discusses the problems and methods employed. Section IV presents the experiments. Section V shows the results obtained and Section VI presents the conclusion and future works.

II. Computer Vision and Object Recognition

Computer Vision is an interdisciplinary area of research that aims at making computers achieve human-like abilities to see and understand images and videos [3]. It includes not only Computer Science but, also, Mathematics, Biology, and Psychology, among other disciplines. The interplay between such areas has led to a great development of computational methods to accurately detect objects in images and videos.

A. Convolutional Neural Network

Convolutional Neural Network (CNN) cite lecun1998 is a type of neural network with state of the art results in image processing and data pattern analysis problems. The biggest advantage this network has over traditional methods is that it needs little manual preprocessing, making it much more viable to use. This is because CNNS automatically produce a feature extractor adjusted to the problem in which it was trained, maximizing its performance and decreasing manual labor. CNN's classic image classification architecture is shown in Figure 1.

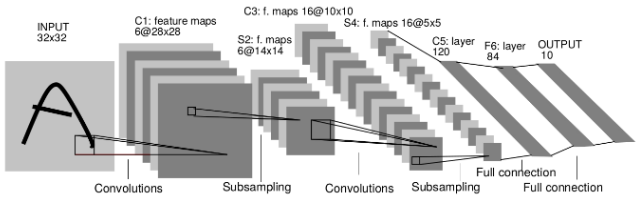


Fig. 1. Architecture of the LeNet5 CNN [4].

B. Transfer Learning

Transfer Learning (TL) is a technique where knowledge used in a particular task is transferred to another similar task. In the area of deep learning is the process of using a CNN model trained in one domain to extract resources in another domain. In this paper, TL was used as a starting point for training the last layers of a CNN. The basic idea of TL is represented in Figure 4.

C. Data Augmentation

Deep learning methods usually require a large amount of data for classification purposes. However, sometimes data is not available in the amount required or classes do not have the same amount of samples. Both problems, small-sized or unbalanced training sets may lead to poor performance. Data augmentation is a useful procedure to enrich the training data, which may reduce overfitting and improve classification performance. Data augmentation has been successfully used in various domains.

In our case, data augmentation is based on generating new images from the original dataset by applying small random transformations to the original images, whilst preserving their labels. These transformations include distortion, zoom, crop, RGB channel shift, among others. For a detailed study of the effect of data augmentation in the performance of CNN image classifiers, see [5].

III. Methods

In this work, the following method was used: the first step was to create a dataset and pre-process raw images. The next step was to train a CNN starting from a pre-trained model (TL). Finally, test images were used to evaluate the classification performance. Figure 2 shows the proposed method flowchart.

A. Data acquisition

For capturing images of vehicles, a Python script was used for web scraping. Images were taken from several websites, including manufacturers' websites and online sale websites, among others. The annotation of make and model was done manually.

B. Pre-processing

Data pre-processing was accomplished in several steps. First, all incompatible image formats (such as GIF, TIFF, PNG, among others) were removed from the initial data collection. Corrupted images due to failures in the web scrapping were also eliminated.

Due to the nature of the problem, there is a very large variability of the images' background within the acquired data. To prevent disruption of the results and cleaning the images, a cropping procedure was done, so as to reduce the background and make the vehicle occupy most of the image. To do this, we used YOLO-v3 (You Only Look Once) [6], pre-trained with the ImageNet dataset [7], to find a bounding box that encloses the vehicle in the image. Though YOLO-v3 is no longer the most accurate, it is very fast and reliable for real-time object identification. Therefore, by using the bounding box coordinates, the image is cropped, as shown in Figure 3. The raw images were not standardized in size, and we used Keras[1] to resize images to the network default input, in this case, to three RGB channels with 224 X 224 pixels.

The next step was to organize the dataset into classes. Images were grouped into classes that correspond to the models of the cars. At this point, the dataset was randomly split into training and test sets, with a ratio of 80% and 20%, respectively.

As a final step, data augmentation was implemented in the algorithm. Since the operations over the images are quite simple, changes were done "on-the-fly", that is, online. Alternatively, one could do this offline, by creating a very large amount of transformed images prior to the training of the algorithm. Online data augmentation requires less storage and guarantees more randomness in the transformations. The transformations used, with respective ranges, are shown in Table I. Moreover, each transformation can be seen individually in Figure 5.

C. Training the Deep Learning Model

Instead of training our make and model classification network from scratch, we choose the Inception-v3 neural network [9] pre-trained on ImageNet as the starting point (TL). The classification layers of the original Inception-v3 networks were discarded, and new classification layers were inserted at the end of the network as following: a DropOut layer with 0.5 probability, a fully connected layer with 1024 neurons and ReLU activation, another DropOut layer with 0.5 probability, and finally a fully connected layer with 50 outputs (number of classes in our classification problem).

The softmax cross-entropy loss function was used to adjust the weights at each iteration by minimizing the classification error between the predicted and expected labels. We employ the Stochastic Gradient Descent (SGD) as the optimizer with Nesterov momentum of 0.9, with a learning rate of 0.01 during the first 15 epochs and 0.002 during the remaining epochs. We

[1]https://keras.io/

Fig. 2. Proposed method.

TABLE I
DATA AUGMENTATION TRANSFORMATIONS. A MORE DETAILED EXPLANATION CAN BE FOUND AT THE KERAS DOCUMENTATION[2].

Arguments	Transformation	Range
Rotation	Rotate the image up to a limit of 40 degrees	40
Width Shift	Move the image horizontally to a fraction of total width	0.2
Height Shift	Move the image vertically to a fraction of total height	0.2
Shear	Turn the rectangular image into a parallelogram-shaped image with a transformation matrix.	0.2
Zoom	Zoom image within the range	0.8-1.0
Channel Shift	Changes the colors schemes given an intensity	30

Fig. 3. (a) Original image, (b) cropped image.

fine tuned the entire network, including the Inception-v3 layers for 30 epochs.

D. Classification and Evaluation

Considering that the problem approached here is supervised learning, the trained CNN model was used for classification so as to find class the image belongs to. Using the data augmentation procedure mentioned before, our dataset was balanced at the level of images by model, that is, all classes have the same amount of samples. Therefore, the evaluation was done using the accuracy in the test set. We computed the overall accuracy and the accuracy for the top-3 ranked classes.

IV. EXPERIMENTS

All data and codes used in this work are freely available for research purposes[3]. The experiments were done with a workstation equiped with two NVIDIA Titan-XP GPUs. Code was developed using Python programming language and Keras.

A. Dataset

There are datasets for car make and model recognition such as: the Cars Dataset [10] of Stanford University and the Vehicle Make and Model Recognition Dataset [11]. However, they have only foreign models, uncommon in Brazil, and one of the motivations of this work is to create an image dataset of cars produced by the Brazilian industry.

Data were collected from various sources, such as vehicle specialized websites, manufacturers' websites, and image repositories in the internet, among others. The initial objective was to classify images in three categories: car make, model and year. However, due to the complexity of the problem, the current version has only the vehicle model.

The UTFPR-CMMD (UTFPR car make and model data set) dataset was created from the list of the 50 best selling cars in Brazil in 2018 for the purposes of brand and car model recognition in still images, the dataset was divided into a proportion of 80% of the training images and 20% of the test images. We provide annotated images for multi-class classification obtained from specialized websites and google images. The full dataset has 24319 images distributed into 10 makes and 50 models. Each class has an average of 500 images. The distribution of the models for each make is shown in Figure 6. The 50 classes contained in the dataset are shown in Figure 7.

V. RESULTS AND ANALYSIS

Two experiments were done: using the original images (with the background) and using the cropped images. This was done to evaluate to what extent the background influences the classification accuracy. For each experiment, the model was trained for 30 epochs. The evolution of accuracy is shown in Figures 8a and 8b.

The final results of the two experiments are shown in Table II. We noticed that the accuracy gain by cropping the cars in the images was very small, thus suggesting that the CNN was able to extract relevant features of the object with very small influence of the background. However, there was a significant improvement comparing the accuracy of the top class and the accuracy based on the top-3 classes. This was due to the fact that the model sometimes confuse models that are quite similar. This can be further analyzed observing the

[3]https://labic.utfpr.edu.br

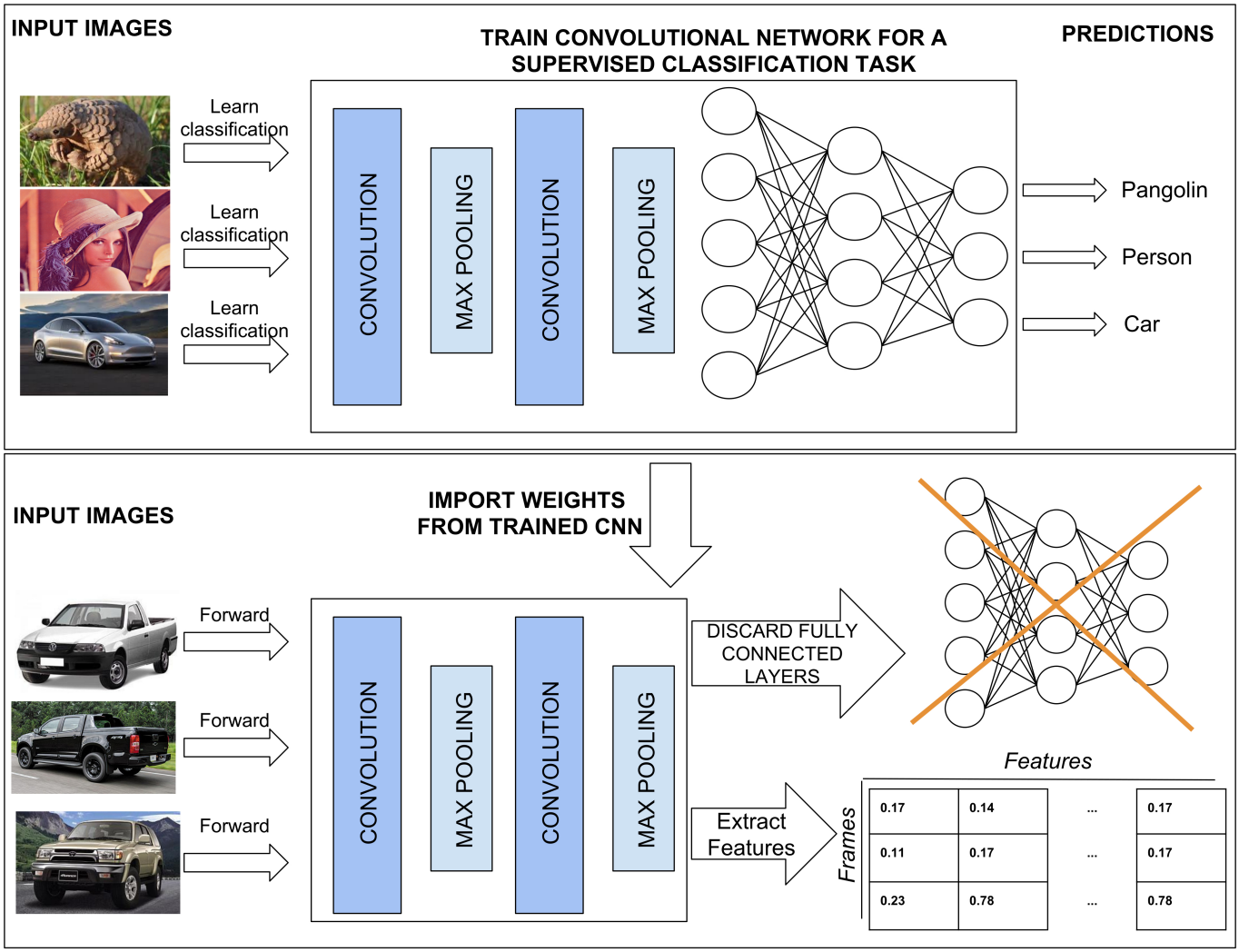


Fig. 4. Basic representation of the Transfer Learning method. Adapted from [8].

confusion matrix of Table III, for the classes where the worst and best results occurred.

Some other experiments were performed in this work, but not reported here, since they did not present significant results. These experiments were the variation of data augmentation parameters and the use of Principal Components Analysis (PCA) to reduce the dimensionality of the features vector.

TABLE II
AVERAGE ACCURACY.

	Normal Images	Cropped Images
Accuracy	79,40%	82,36%
Top-3 accuracy	95,43%	94,87%

VI. CONCLUSION

Car make and model recognition in images or videos is a non-trivial problem since some models show no apparent differences from each other when seen from certain view angles. This work presented a slightly different approach from those in the literature, where the images are only of a specific angle and low diversity.

TABLE III
CONFUSION MATRIX FOR WORST AND BEST RESULTS.

	HB20	HB20S
HB20	138	144
HB20S	165	98

	Amarok	Captur
Amarok	276	0
Captur	0	257

The proposed model was able to achieve good results, provided the dataset had a large diversity of images. The use of data augmentation was important to balance the classes, leading the classifier to achieve a higher accuracy.

Actually, the most challenging part of this work was building and labeling the dataset. Therefore, this is a significant contribution to other researchers interested in car make and model recognition.

Future work should focus on increasing the performance of the classifier, using, for instance, a Learning Vector Quantization (LVQ) approach. Another way to improve the results is

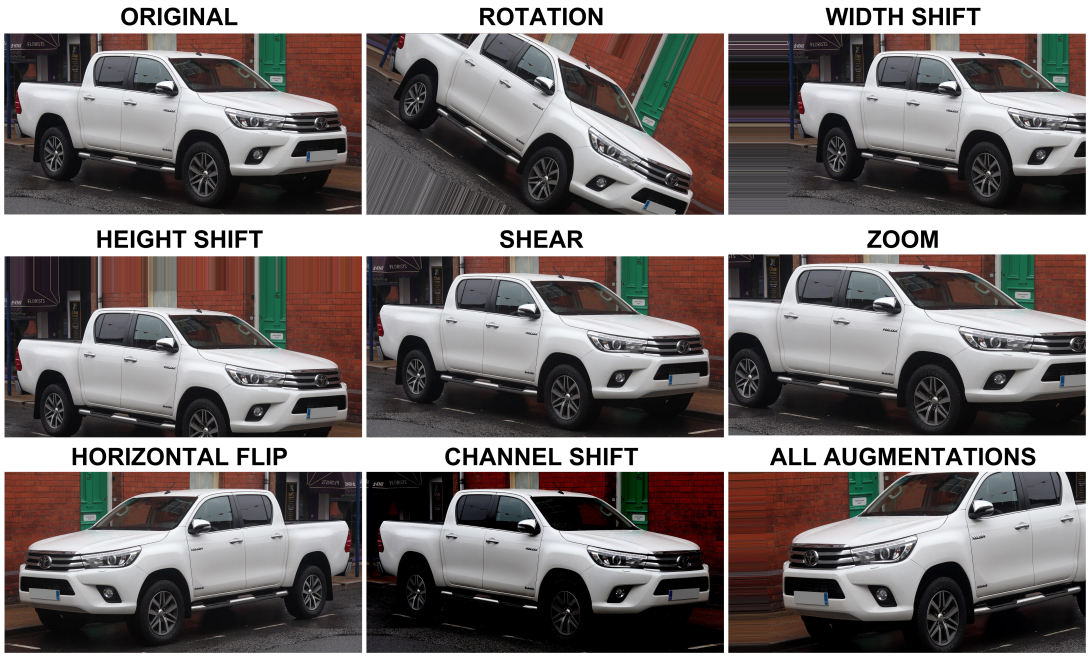


Fig. 5. Types of data augmentation used in this work.

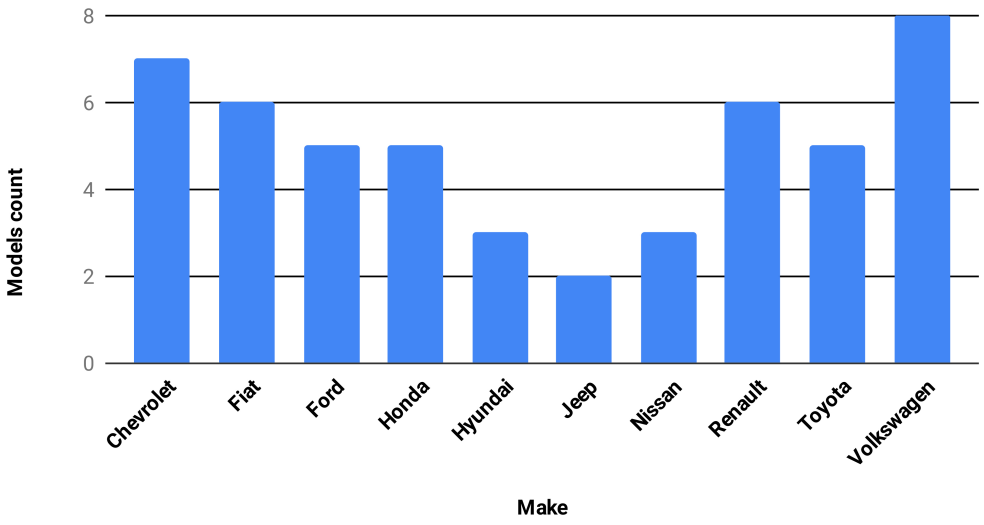


Fig. 6. Distribution of models for different makes.

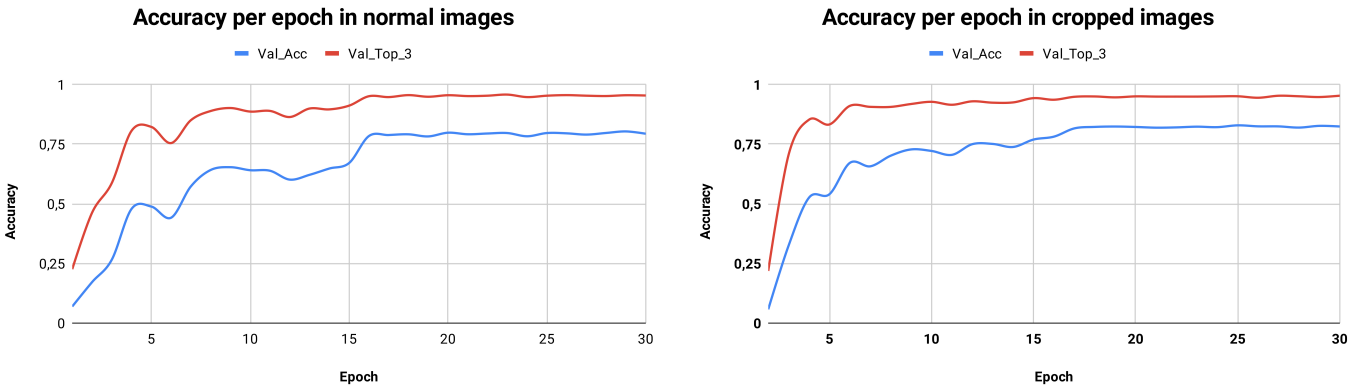
to hybridize the CNN with other methods, such as template matching for locating the brand logo, when possible. Regarding the dataset, it is intended to extend with more makes and models. Possibly, the developed approach will evolve so as to be integrated into public security and/or traffic control system.

ACKNOWLEDGEMENTS

L.A. Albini thanks CNPq for the PIBIC scholarship, M Gutoski thanks CNPQ for the scholarship. H.S.Lopes thanks to CNPq for the research grants no. 311778/2016-0 and 423872/2016-8, and Fundação Araucária for the financial support by means of PRONEX 042/2018. All authors thanks NVIDIA for the donation of Titan-Xp GPU boards.

Make	Model	Make	Model	Make	Model
Chevrolet	Ônix	Honda	HR-V	Toyota	Corolla
	Prisma		Fit		Etios
	S10		Civic		Hilux
	Tracker		WR-V		Etios Sedan
	Spin		City		SW4
	Cruze	Hyundai	HB20	Volkswagen	Polo
	Cobalt		Creta		Gol
Fiat	Strada		HB20S		Saveiro
	Argo	Jeep	Compass		Virtus
	Toro		Renegade		Amarok
	Mobi	Nissan	Kicks		Fox
	Cronos		Versa		Voyage
	Grand Siena		March		UP
Ford	Ka	Renault	Kwid		
	Ka Sedan		Sandero		
	Ecosport		Logan		
	Fiesta		Captur		
	Ranger		Duster		
			Duster Oroch		

Fig. 7. Classes of the UTFPR-CMMD dataset.



(a) Original image

(b) Percentage Of colors

Fig. 8. Accuracy per epoch using cropped images.

REFERENCES

[1] S. Agarwal, J. O. du Terrail, and F. Jurie, "Recent advances in object detection in the age of deep convolutional neural networks," *CoRR abs/1809.03193*, 2018.

[2] J. Boyle and J. Ferryman, "Vehicle subtype, make and model classification from side profile video," in *Proc. 12th IEEE International Conference on Advanced Video and Signal based Surveillance (AVSS2015)*. IEEE Press, 2015, pp. 1–6.

[3] C. Steger, M. Ulrich, and C. Wiedemann, *Machine Vision Algorithms and Applications*, 2nd ed. Weinheim: Wiley VCH, 2018.

[4] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.
works," in *Proc. XIII Brazilian Congress on Computational Intelligence (CBIC)*. Niterói, RJ: ABRICOM, 2017.

[6] J. Redmon and A. Farhadi, "YOLO-v3: An incremental improvement," *arXiv:1804.02767v1*, 2018.

[5] N. M. R. Aquino, M. Gutoski, L. T. Hattori, and H. S. Lopes, "The effect of data augmentation on the performance of convolutional neural net-

[7] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.

[8] M. Gutoski, "Learning and transfer of feature extractors for automatic anomaly detection in surveillance videos," MSc. Dissertation, Graduate Program in Engineering and Computer Science, Federal University of Technology Paraná – UTFPR, 2018.

[9] S. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception architecture for computer vision," *CrXiv:1512.00567v3*, 2015.

[10] J. Krause, M. Stark, and J. D. andLi Fei-Fei, "3D object representations for fine-grained categorization," *Proc. 4th IEEE Workshop on 3D Representation and Recognition*, 2013.

[11] F. Tafazzoli, K. Nishiyama, and H. Frigui, "A large and diverse dataset for improved vehicle make and model recognition," *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.