Journal of Intelligent & Robotic Systems (2021) 101: 61
https://doi.org/10.1007/s10846-021-01334-0

On Robustness of Robotic and Autonomous Systems Perception

An Assessment of Image Distortion on State-of-the-art Robotic Vision Model

Cristiano Rafael Steffens[1] · **Lucas Ricardo Vieira Messias**[1] · **Paulo Jorge Lilles Drews-Jr**[1] · **Silvia Silva da Costa Botelho**[1]

Received: 6 October 2020 / Accepted: 27 January 2021 / Published online: 3 March 2021

Abstract

We propose an evaluation framework that emulates poor image exposure conditions, low-range image sensors, lossy compression, as well as noise types which are common in robot vision. We present a rigorous evaluation of the robustness of several high-level image recognition models and investigate their performance under distinct image distortions. On one hand, F1 score shows that the majority of CNN models are slightly affected by mild exposure, strong compression, and Poisson Noise. On the other hand, there is a large decrease in precision and accuracy in extreme misexposure, impulse noise, or signal-dependent noise. Using the proposed framework, we obtain a detailed evaluation of a variety of traditional image distortions, typically found in robotics and automated systems pipelines, provides insights and guidance for further development. We propose a pipeline-based approach to mitigate the adverse effects of image distortions by including an image pre-processing step which intends to estimate the proper exposure and reduce noise artifacts. Moreover, we explore the impacts of the image distortions on the segmentation task, a task that plays a primary role in autonomous navigation, obstacle avoidance, object picking and other robotics tasks.

Keywords Image enhancement · Image restoration · Deep neural networks

1 Introduction

Vision has become a catalyst in the implementation of robotics automated systems that rely on environment perception. Among the practical applications of this hardware-software associations are biometric surveillance systems [23], automated visual inspections [37, 49], object tracking [45, 58, 69], environment mapping [5], the domestic and assistive robots [22, 39], field robotics [62, 63], and autonomous cars [2, 26]. Moreover, vision-based perception has also proved valuable in diverse robotics and autonomous systems, being used for visual servoing [65], obstacle detection and avoidance in unmanned vehicles [6, 7, 12, 35], localization and mapping [15], navigation [56], distance estimation [11], loop closure [41], and picking and grasping robots [24, 33, 34, 40]. The vision part in these tasks, as well as in other machine vision tasks, is often provided by convolutional neural networks (CNNs). However, researchers have shown CNNs to be vulnerable to image distortion and manipulation, which may undermine the reliability of CNN-based systems in autonomous, vision-based, and robotic applications.

Amidst the recent development of Convolutional Neural Networks (CNN), such as DenseNet [21], Inception-v3 [53], Inception-v4 and Inception-ResNet-v2 [52], MobileNetV1 [19], MobileNet-v2 [46], NASNet [71], NASNetMobile [71], ResNet [17] ResNet-v2 [18], ResNeXt [64], VGG [48], and Xception [4], it becomes evident that the advancements accomplished within the classifications based on Deep Learning (DL) have reached the level of near-human accuracy in numerous datasets.

✉ Cristiano Rafael Steffens
cristianosteffens@furg.br

Lucas Ricardo Vieira Messias
lucasmessias@furg.br

Paulo Jorge Lilles Drews-Jr
paulodrews@furg.br

Silvia Silva da Costa Botelho
silviacb@furg.br

[1] NAUTEC - Intelligent Robotics and Automation Group, FURG - Universidade Federal do Rio Grande, Rio Grande, Brazil

When it comes to the improvement of machine learning-based models, vast and assorted datasets are fundamental. Many learning-based image recognition systems advance in order to boost its precision within one of the following datasets: Imagenet ILSVRC15 [44], MS-COCO [32], Open Images [29], AVA [14, 30], KITTI [13], CIFAR [28], and PASCAL VOC [9]. Although nowadays the quantification and generalization of a prepared model's efficacy is customarily evaluated by its aptitude on a given test set. The aforementioned datasets are the baseline for the development of a multitude of the current classification models, contributing with a true touchstone for the comparison and assessment of distinct models.

The effectiveness of such evaluation, however, is a reason of concern for researchers as DL-based systems expanded to be the standard in perception for robotics. This is evidenced in how Osherov et al. [38] investigated the consequences of partial occlusion in the performance of a trained image classification network model. Recht et al. [42] also suggested that the over-adjustment of DL-based classifiers to increase the precision of predictions in test sets may render the system incapable of generalizing to images marginally more difficult than the ones offered in the initial test sets.

Another source of worry for researchers over the robustness of DL based classifiers is what is called adversarial perturbations - minor alterations that are intended to change the classifier's estimations. Szegedy et al. [54] points out that deep networks are surprisingly vulnerable to such perturbations. Even though they are hardly perceptible to the human eye, adversarial perturbations are enough to cause a deep network's decision to switch. This has been observed in later studies [10, 25], which have brought on similar results for classification and segmentation tasks.

The accuracy of image recognition models, despite the thorough exploration, still presents a limitation that often goes unnoticed, and that difficulty is the question if CNNs are ready to classify images that do not display proper exposure, noise levels, or over-compression. Whereas such circumstances prove to be common in any vision-based pipeline, their effects on the final predictions' performance have not been inspected meticulously. The research on this topic has become increasingly relevant since the accident with an autonomous car in development by Uber Technologies Inc., which has been thoughtfully discussed by Kohli et al. [26]. High contrast and overexposure due to shadows and strong sunlight has also been reported as critical for self-driving cars by Zhang et al. [69] and Wang et al. [61].

This paper aims to both offer a methodology and a theoretical framework to support the investigation on the impact of image distortions on image recognition models and to evaluate the robustness of an image recognition model's using a comprehensive set of images against common distortions. We extend a previous publication from Steffens et al. [51] which provided a short insight on the impacts of image distortions in image recognition models. In this paper, we expand the discussion on how damaged images impact a series of classifier performance metrics. We also include a set of new image recognition models which have obtained outstanding results over the last years. Furthermore, we also deepen the discussion on the main novelties introduced by each of these models and how they can impact their resilience under non-ideal image conditions. We believe the methodology can be consistently applied to validate a wide range of vision-based applications.

2 Background Theory

In order to assess the resilience of diverse DL-image recognition models, we propose the use of synthetically generated sets of images in conditions of over-compression, noise, overexposure, and underexposure. The classification models were used as originally proposed, *i. e.* they were used with identical sets of weights, input shapes, and pooling layers provided by their authors. The models were previously prepared in order to cope with the ImageNet ILSRVC Challenge [44].

The ILSRVC Challenge rules establish that each recognition model should output a list of, at most, five categories arranged by confidence for each image. The virtue of such classification is judged by the label that corresponds the best to the ground truth label given to the image, which permits an algorithm to identify more than one object within an image without being penalized if one identified object was there but not included in the ground truth.

2.1 Distortions

Figure 1 shows the effects of diverse distortions applied to the images in order to evaluate the robustness of distinct image classifiers towards bad exposure and noise. The details about the image distortions and how they occur in real world applications is are discussed bellow. We have released a Python implementation of each of the distortions described in this section on Github, in the https://git.io/JUgIz repository.

2.1.1 JPEG Over-compression

JPEG (Joint Photographic Experts Group) compression usually results in block artifact and high-frequency aliasing, also known as mosquito noise, due to its mild cosine transformation (DCT), which is block-wise [55]. The tests were proceeded with input images in a high JPEG compression ratio, with the quality $Q = 15$, a loss of detail that is

Fig. 1 Example of the image distortions applied in the evaluation of the robustness of CNN-based image recognition models: **a** Original image, **b** JPEG over-compression, **b** Additive White Gaussian Noise, **d** Salt & Pepper Noise, **e** Speckle Noise, **f** Poisson Noise, **g** Gamma $\frac{1}{2}$, **h** Gamma $\frac{1}{4}$, **i** Gamma $\frac{1}{8}$, **j** Gamma 2, **k** Gamma 4, **l** Gamma 8, **m** Truncation Q_1, and **n** Truncation Q_3

common in any machine vision pipeline with unstable or limited bandwidth.

2.1.2 Additive White Gaussian Noise (AWGN)

AWGN is a noise added randomly to the image with normal distribution as the probability density function [27]. A standard deviation value of $\sigma = 23.55$ was, in this case, used to cause a decline in the quality of the image. Additive White Gaussian Noise models are widely used in the literature provide a coarse approximation of real sensor noise [1].

2.1.3 Salt & Pepper Noise (S&P)

S&P is an impulse noise with the same salt-and-pepper probability on a pixel, where the salt is a bright pixel (with 255 pixel value) and the pepper is a dark pixel (with 0 pixel value), with a probability per pixel of P [47]. Within the scope of the tests hereby described, P was valued in 0.3. In real applications, this kind of distortion is most commonly related to a malfunctioning of a camera's sensor cell, which generates dead pixels.

2.1.4 Speckle Noise

Speckle noise is originated from coherent processing of back-scattered signals from multiple distributed points [57], as a product of environmental conditions on the imaging sensor during the acquisition of an image. The Speckle Noise of an image I can be expressed as $\hat{I} = I + (n \times I)$, where n is a uniform noise with mean $\mu = 0$ and variance $\sigma^2 = 1$. This type of noise is commonplace in medical images, SAR (Synthetic Aperture Radar) images, and active Radar images [36].

2.1.5 Poisson Noise

Poisson Noise also known as shot noise, this noise occurs due to fluctuations in electrical currents that are due to chance arrivals of electrons to an anode [59]. It is correlated with the pixel values in the input image. Poisson noise is often assumed to be inherent to virtually all images on some level [67].

2.1.6 Gamma Power Transformation

Gamma Power Transformation constitutes a nonlinear operation utilized to encode or decode luminance values in image systems [55]. One of its usages is to adjust or compensate certain luminance levels in images. The images are emulated under and overexposed as $\hat{I} = I^{\gamma}$. The power transformation is followed by a min-max normalization, which is necessary in order to adjust the pixel values into a valid representation range. This, results in data loss for either light or dark regions, depending on $\gamma < 1$ or $\gamma > 1$, respectively, therefore emulating the typical poor exposure conditions in mobile robotics.

2.1.7 Quantile-based Truncation

Quantile-based Truncation is used to mimic underexposure when pixels are truncate on Q_1, and overexposure when the pixels are truncate on Q_3, imitating the quality of the image captured by the cheaper imaging sensors with low dynamic range - considering that Q_1 and Q_3 are the first and third quartile of an image's pixel's distribution.

2.2 Image-Recognition and Segmentation Models

This paper proposes a methodology to evaluate the effects of image degradation on image recognition tasks. The proposal is applied to evaluate the adjustment of CNN based classification models for the ImageNet Large Scale Visual Recognition Challenge (ILSVRC). To classify objects presented in an image is the purpose of Image classification models. Over the years, Deep Learning grew into the standard way of solving image classification problems. The ILSRVC Challenge made it so that the models' architecture shifted to attain the best

classification accuracy, within the numbers of parameters in the network.

2.2.1 VGG

Proposed by Simonyan et al. [48], VGG has achieved both first and second places in the ILSVRC-2014, mainly because, as it has been proven, the increase on the depth of the network (*i.e.* stack more layers) in combination with small (3×3) convolution filters, ends up meaning great improvement over previously tried methods. While VGG requires a big amount of computational resources because of its large width of convolutional layers, VGG secured its place as one of the most widely used systems for feature extraction in perceptual, style, and contextual loss applications. It has been integrated into the training strategies of various image-to-image translation deep learning models.

2.2.2 ResNet

Proposed by He et al. [17], the ResNet model has obtained first place in the ILSVRC-2015. The authors explicitly reformulate the layers as learning residual functions regarding the layer inputs, instead of learning unreferenced functions (residual module). By doing so, the model is able to avoid both the vanishing gradient problem, as well as the degradation problem in optimization. In terms of structure, the model is composed mostly of 3×3 convolutions and average pooling layers.

2.2.3 Inception-v3

Proposed by Szegedy et al. [53], the Inception-v3 model presents factorized convolutions and aggressive regularization, scaling up the network's efficient and improving accuracy. The Inception module uses convolutions of different sizes to capture details at varied scales (5×5, 3×3, 1×1).

2.2.4 Inception-ResNet-v2

Proposed by Szegedy et al. [52], the Inception-ResNet-v2 shows that combining an inception's uniform simplified architecture with residual connections and more inception modules than [53], accelerates the training and achieve betters accuracy results.

2.2.5 DenseNet

Proposed by Huang et al. [21], DenseNet is a model based on [17] where each layer obtains additional inputs from all preceding layers and passes on its own feature-maps to all subsequent layers instead of the element-wise addition between the input and the output of a layer.

2.2.6 Xception

Proposed by Chollet et al. [4], the Xception network is inspired by the inception's architectures having skip connections and modified depth-wise separable convolutions as an improvement.

2.2.7 MobileNetV1

Proposed by Howard et al. [19], MobileNetV1 is a CNN with depthwise separable convolutions between the regular convolutions layers. Hence, the parameters and multiply-add operations are considerably reduced, which is suitable for mobile devices, or any devices with low computational power.

2.2.8 MobileNetV2

Proposed by Sandler et al. [46], MobileNetV2 network has a residual block with a stride of 1 and a block with a stride of 2 for downsizing, outperforming [19].

2.2.9 NASNet

Proposed by Zoph et al. [71], NASNet is a model that utilizes the information acquired on a small dataset over a big one searching for the best convolutional layer on the first one. The authors also propose `ScheduledDropPath` regularization technique, which significantly improves the model's generalization.

2.2.10 Mask-RCNN

Proposed by He et al. [16], Mask-RCNN is a model for instance segmentation (*i.e.* to find instances of a countable object in the scene). Its goal is to distinguish each instance of each object within the image at the pixel level. This model is based on FPN (Feature Pyramid Networks for Object Detection) by [31] which improves its hability for detecting objects at different scales. Mask-RCNN consists of two stages. First, it generates proposals about the regions where there might be an object. Second, it predicts the class of the object, creates a bounding box and generates a mask in pixel level of the object based on the first stage proposal. The two stages are connected to the backbone structure.

Table 1 provides details concerning release dates, network size, input image resolution, and Top-1 validation accuracy attained by every model, in accordance with

Table 1 Classification Models Considered in the experiments. Top-1 validation accuracy according to the official reports

Model	Year	Size	Parameters	Top-1	Resolution
VGG-16 [48]	2014	528 MB	138,357,544	0,71	224
ResNet50 [17]	2016	98 MB	25,636,712	0,75	224
Inception-v3 [53]	2016	92 MB	23,851,784	0,78	299
Inception-ResNet-v2 [52]	2017	215 MB	55,873,736	0,80	299
DenseNet201 [21]	2017	80 MB	20,242,984	0,77	224
Xception [4]	2017	88 MB	22,910,480	0,79	299
MobileNetV1 [19]	2017	16 MB	4,253,864	0,70	224
MobileNetV2 [46]	2018	14 MB	3,538,984	0,71	224
NASNetLarge [71]	2018	343 MB	88,949,818	0,83	331
NASNetMobile [71]	2018	23 MB	5,326,716	0,74	224

the official reports.[1] Due to pre-processing, deep learning framework optimization, and float points precision utilized during inference, the actual accuracy may vary. As to avoid the interference of these variables, we rerun the inference on the original validation set under the same conditions as all other distorted images. Nonetheless, the evaluation methodology can also be applied to any model and dataset for image-recognition or segmentation tasks.

3 Problem Statement and Methodology

Autonomous robotics often rely on visual perception of the environment to perform their tasks. Visual perception includes instance segmentation, semantic image segmentation, pan-optic segmentation, depth-estimation and classification. Vision based feedback has been extensively used for interaction for mapping and navigation, human robot interaction, and robot environment interaction.

3.1 Problem

The robustness of these vision based models against commonly occurring image distortions and noises has not yet been properly addressed. The development of machine vision systems often focuses only on maximizing the accuracy of the model on one set of samples, disregarding conditions which can significantly impact autonomous systems in real applications.

[1] Updated state-of-the-art results are available at https://paperswithcode.com/sota/image-classification-on-imagenet. Challenge winners for each edition can be found at http://www.image-net.org/challenges/LSVRC/2017/results, http://www.image-net.org/challenges/LSVRC/2016/results, http://www.image-net.org/challenges/LSVRC/2015/results, and http://www.image-net.org/challenges/LSVRC/2014/results.

Considering that the majority of modern autonomous systems rely on CNNs to extract information from the environment, it is necessary to establish a practical approach to asses the performance degradation linked to basic image distortions. Furthermore, we notice that most most CNN models used for classification and segmentation nowadays share the same building blocks.

Our research intends to answer the following questions: *i.* What are the impacts of image distortions on popular CNN architectures?, *ii.* Which distortions are the most critical for robot vision?, *iii.* Can we build better pipelines to make robot vision less vulnerable towards ill exposure and noise?, and *iv.* Do the same distortions impact different robot vision tasks?

3.2 Methodology

First, we assess the impacts of common image distortions on popular CNN architectures. This evaluation is performed on object recognition, considering this field has been fully explored over the last decade, with many recognition models achieving near-human accuracy. Moreover, these models also set the directions in the development of other vision tasks and are, therefore, a good measure of the overall performance of vision perception for robotics.

The assessment of the robustness of relevant image recognition models in the literature is as follows. A detailed evaluation of the robustness of state-of-the-art image recognition networks for common distortion of images was performed, with metrics calculated on the ILSRVC ImageNet Challenge validation sub-set.

1. The images are individually inserted through the Python Scikit-Image [60] library within the parameters of 8-bit unsigned integer data type, where all files are stored as compressed JPEG format, with variations in image size and ratio.

2. The input image goes through a distortion process, following the description in Section 2.1. While noise is usually linked to the sensor and electronics used for image acquisition, compression is often a mandatory requirement to preserve the frame rate and storage space. Badly exposed images are usually linked to sensors with a limited dynamic range or poor selection of lens aperture, exposure time, or gain, even in auto-adjusted acquisition setups.
3. The images were resized and cropped to the constraints accepted by the model, moment when a first-order spline interpolation is used. Gaussian filter at $\sigma = \frac{s-1}{2}$ was used as an anti-aliasing method for downscaling the images, s being the scaling factor. If the model required so, further specific image transformations were done to adapt the input and data representation.
4. The image is finally ready to be inserted as input to the system, and, once the inference proceeds, the results are then stored for further evaluation.
5. This evaluation judges a few of the most popular evaluation metrics: Top-1 Accuracy, Top-5 Accuracy, and F1-Score. Taking into consideration the number of true instances for each label, a weighted average with only Top-1 results is used for Precision, Recall, and F1-Score.

Then, taking in account the results obtained in the first set of experiments, we explore the relationship between accuracy and noise severity. This assessment is performed considering the noise types that have shown to be critical for all recognition models. Thus, we include a study with the simplest model and with the distortions (using variable parameters).

After that, we investigate an alternative robot vision pipeline. In order to minimize the undesirable impacts of noise and miss-exposure, we modify the traditional pipeline by introducing an image restoration step, which operates in the RGB colorspace, after the image has already been compressed and transmitted. The restoration takes place immediately before the object recognition algorithm so that the pipeline can be easily applied to other vision tasks without requiring further adaptations or hardware customization. We add two image pre-processing steps to deal with ill-exposure and noise.

Finally, in order to investigate if the findings from the assessment on the object recognition task also hold for distinct applications of perception in robotics, we investigate how noise and inadequate exposure affect instance segmentation. Segmentation models are frequently applied for tasks such as object picking, obstacle detection and avoidance, autonomous navigation, and human-robot interaction. These systems often share the same basic building blocks that have show successful with CNN based object recognition models [7, 15, 24, 34, 35, 41, 56, 69]. As these models rely on the same set of techniques and popular architectures, we expect them to show the same strengths and weaknesses as the object recognition CNN models.

4 Experiments and Results

This Section is organized in four subsections according to the objectives outlined in the methodology. The results shown in Section 4.1, Section 4.2, and Section 4.3 take into account the results obtained on the Imagenet ILSRVC Validation subset. This subset consists of 50,000 images from 1000 distinct categories. The samples are evenly distributed so that each category presents exactly 50 images. This limits the false-negative (FN) predictions per class to the number of samples in it. It is necessary to stress, that the Top-1 Accuracy rendered by the appraised models are not identical to the results in the official reports. The aforementioned distinction arises from the software used for image processing, image resizing interpolation strategies, image cropping strategies, or perhaps even float-point precision. Nonetheless, the procedure is the same for all evaluated models. Section 4.4 exemplifies the issues with image distortions on the instance segmentation task, which aims to distinguish each instance of each object within the image. Segmentation is an ordinary function for intelligent robotics and automation.

4.1 Impacts of Image Damage on Object Recognition

For each evaluated model, we present tabulated results which include Top-1, and Top-5 accuracy, as well as the F1-Score. We also discuss the box-plots for false-positives (FP) and false-negatives (FN), which provide insight into the effects the image distortion has on each class. Table 2 summarizes the results for all models evaluated. We highlight each result in the presented tables according to the following strategy. Results for undamaged images are shown in black. Conditions that worsened the accuracy by up to 10% are marked in green (low impact). Conditions that worsened the performance of the model by any value between 10% and 30% are shown in orange (moderate impact). Conditions that worsened the network outcomes by more than 30% are shown in red (critical impact). This scale allows for fast visualization of the models' robustness in the light of its original performance.

The results achieved by the VGG-16 [48] model, when confronted with a myriad of distortions show that the model performed with noticeable robustness under moderate ill exposure ($\gamma = [\frac{1}{2}; \frac{1}{4}; 2; 4]$, Overexposure Q_3, Underexposure Q_1), heavy compression, Gaussian

Table 2 Impacts of the input image distortion on the object recognition task

	VGG-16			Resnet			Inception-v3			Inception Resnet-v2			DenseNet		
Metric	Top-1	Top-5	F1-Score	Top-1	Top-5	F1-Score	Top-1	Top-5	F1-Score	Top-1	Top-5	F1-Score	Top-1	Top-5	F1-Score
Original	0.612	0.838	0.609	0.668	0.870	0.666	0.747	0.920	0.744	0.773	0.936	0.770	0.663	0.870	0.664
Gamma 1/2	0.584	0.817	0.582	0.634	0.848	0.633	0.727	0.908	0.724	0.755	0.926	0.752	0.602	0.830	0.608
Gamma 1/4	0.455	0.707	0.464	0.501	0.745	0.509	0.645	0.854	0.647	0.683	0.879	0.685	0.445	0.696	0.465
Gamma 1/8	0.236	0.452	0.252	0.280	0.503	0.302	0.469	0.703	0.487	0.516	0.743	0.533	0.222	0.429	0.243
Gamma 2	0.566	0.800	0.564	0.623	0.838	0.621	0.719	0.905	0.716	0.746	0.920	0.742	0.625	0.841	0.626
Gamma 4	0.401	0.635	0.408	0.459	0.693	0.467	0.591	0.809	0.593	0.626	0.836	0.626	0.458	0.685	0.468
Gamma 8	0.175	0.334	0.192	0.217	0.390	0.235	0.342	0.552	0.361	0.385	0.595	0.402	0.224	0.398	0.242
Q1 trunc.	0.541	0.780	0.539	0.593	0.814	0.592	0.713	0.900	0.709	0.737	0.916	0.734	0.642	0.854	0.641
Q3 trunc.	0.548	0.780	0.546	0.603	0.816	0.602	0.721	0.903	0.718	0.750	0.922	0.747	0.642	0.854	0.642
JPEG Q=15	0.538	0.775	0.535	0.568	0.795	0.568	0.671	0.873	0.668	0.707	0.895	0.704	0.588	0.818	0.592
Gauss	0.508	0.743	0.507	0.534	0.765	0.533	0.687	0.882	0.684	0.716	0.900	0.713	0.588	0.814	0.592
Poisson	0.586	0.816	0.582	0.626	0.841	0.623	0.732	0.911	0.729	0.759	0.927	0.756	0.651	0.862	0.652
S&P	0.143	0.290	0.155	0.126	0.260	0.137	0.362	0.586	0.374	0.405	0.623	0.418	0.191	0.372	0.213
Speckle	0.081	0.182	0.089	0.069	0.156	0.078	0.261	0.459	0.272	0.311	0.517	0.329	0.145	0.299	0.162
	Xception			Mobilenet			Mobilenet V2			NASNetLarge			NASNetMobile		
Metric	Top-1	Top-5	F1-Score	Top-1	Top-5	F1-Score	Top-1	Top-5	F1-Score	Top-1	Top-5	F1-Score	Top-1	Top-5	F1-Score
Original	0.763	0.929	0.760	0.657	0.863	0.655	0.600	0.827	0.603	0.806	0.951	0.803	0.693	0.888	0.689
Gamma 1/2	0.746	0.920	0.743	0.623	0.840	0.624	0.497	0.751	0.511	0.794	0.946	0.791	0.663	0.868	0.661
Gamma 1/4	0.662	0.869	0.665	0.504	0.746	0.515	0.295	0.535	0.316	0.745	0.919	0.743	0.551	0.780	0.558
Gamma 1/8	0.452	0.452	0.252	0.285	0.513	0.313	0.130	0.285	0.139	0.612	0.828	0.620	0.324	0.542	0.349
Gamma 2	0.734	0.915	0.731	0.614	0.832	0.612	0.564	0.795	0.566	0.787	0.941	0.784	0.652	0.858	0.649
Gamma 4	0.612	0.828	0.612	0.452	0.686	0.461	0.376	0.605	0.389	0.691	0.882	0.690	0.486	0.716	0.491
Gamma 8	0.370	0.586	0.386	0.218	0.396	0.238	0.157	0.305	0.174	0.467	0.681	0.480	0.239	0.416	0.257
Q1 trunc.	0.729	0.910	0.726	0.612	0.831	0.610	0.577	0.806	0.577	0.777	0.936	0.774	0.647	0.857	0.643
Q3 trunc.	0.734	0.914	0.731	0.621	0.839	0.619	0.587	0.816	0.588	0.785	0.940	0.782	0.659	0.863	0.655
JPEG Q=15	0.688	0.884	0.686	0.567	0.798	0.568	0.497	0.743	0.504	0.500	1.000	0.500	0.620	0.837	0.618
Gauss	0.701	0.893	0.697	0.544	0.775	0.544	0.436	0.677	0.447	0.765	0.930	0.761	0.607	0.826	0.604
Poisson	0.747	0.920	0.744	0.627	0.843	0.625	0.567	0.801	0.571	0.796	0.946	0.792	0.671	0.872	0.667
S&P	0.382	0.610	0.392	0.189	0.363	0.196	0.070	0.168	0.075	0.527	0.746	0.531	0.243	0.430	0.254
Speckle	0.261	0.459	0.272	0.093	0.205	0.101	0.041	0.102	0.045	0.423	0.638	0.434	0.167	0.315	0.183

and Poisson Noise, when it came to the classification accuracy. Salt and Pepper and Speckle Noise, however, culminated in an intense decrease in the classifier's accuracy and precision. Considering the F1-Score, given by $F1 = 2 \times \frac{Precision \times Top1}{Precision + Top1}$, which provides a good assessment of both false positives and false negatives, we can quickly identify the distortions that have a critical impact. When the predictions of the VGG-16 model on the original images are compared to predictions on damaged images, we notice that Gamma transformation with $\gamma = [\frac{1}{8}; 8]$, S&P and Speckle Noise, more than halved the performance of the classifier.

Resnet shows to be robust against mild exposure variations ($\gamma = [\frac{1}{2}; 2]$) and Poisson Noise. Pixel value truncation, otherwise, shows to have a higher impact dropping Top-1 accuracy levels from 0.668 to 0.603, for truncation in the bright part, and 0.593, for truncation in the darker pixels. Compression, coarse miss-exposure, Gaussian Noise, impulse Noise, and Speckle Noise also result in a dramatic decrease in all metrics considered for this evaluation. We notice the same patterns observed for the VGG-16 network. Mild misexposure generated through power transformation $\gamma = [\frac{1}{2}, 2]$ and quantile-wise truncation are the distortions with the least impact the classifiers' performance. Poisson Noise and over-compression also have a mild impact on the results. Signal independent noise results in most images being miss-classified in a few categories, as shown by the distribution of FPs and FNs. Harsh power transformation with $\gamma = \frac{1}{4}, \frac{1}{8}, 4, 8$ also result in expressive impacts.

As for Inception-v3 [53] image recognition model, we notice that it performs close to the official reports in our setup for the image without any distortion. Taking into account the F1-Score, which provides a good visualization of both false positives and false negatives, we observe that this model is less affected by image distortions than the aforementioned VGG-16 and Resnet50. The worst performance is observed when the images include Speckle Noise and S&P Noise, followed by images severely damaged due to synthetic miss-exposure generated using Gamma power transformation.

Inception-ResNet-v2's achieved a performance that can (25%), and Poisson Noise. Lossy image compression and gross Gaussian Noise, show to have a moderate impact. The robustness towards Gaussian Noise is expected, once it is a data augmentation strategy often used during model training. Gross miss-exposure generated through power transformations with $\gamma = [\frac{1}{4}; \frac{1}{8}; 4; 8]$, however, has shown to have expressive impact in the classification accuracy. The metrics show a mirrored effect, displaying similar results for both dark and bright images. None of the above, however, seems to affect classification accuracy as much as the impulse noise and Speckle Noise.

The results concerning Xception's network [4], show that this network is also susceptible to common image distortions. This model obtains a good level of robustness when submitted to Gamma power transformations with $\gamma = [\frac{1}{2}; \frac{1}{4}; 2; 4]$, quantile-based truncation with both Q_1 and Q_3, Gaussian and Poisson Noise. Distortions more intense than $\gamma = [\frac{1}{8}; 8]$, Speckle Noise, and S&P Noise, however, have a non-negligible negative impact on the accuracy.

The MobileNetV1 [19] model is influenced by distinct distortions. The network is one of the smallest among all models considered in this study, using only 4,253,864 trainable parameters against 138,357,544 in VGG-16. It is necessary to remark that, even on images with no applied distortion, the model displays lower accuracy than the level reported in the official reports (summarized in Tab. 1). Not unlike the major part of the evaluated models, S&P and Speckle Noise culminate in significant considerable decay. This information becomes particularly relevant when confronted with the fact that these distortions are usually linked to the first steps of any vision pipeline.

The results obtained by MobileNetV2 [46] indicate that this model has a lower performance than its predecessor. Only the original images, $\gamma = 2$, quantile-wise truncated images, and images affected by Poisson Noise resulted in Top-1 accuracy higher than 0.5. Image compression, power transformation, and noise resulted in expressive accuracy drops. As in other evaluated models, S&P and Speckle Noise show to have a significant impact on the image recognition quality. In terms of trainable parameters, MobileNetV2 is the smallest image classification model considered in the present study.

Finally, Table 2 also shows the results for NASNet models. These models use cells found through optimized Network Architecture Search (NAS) where distinct network components are evaluated in order to design the best performing network architecture. NASNetLarge [71] takes an input image with $331 \times 331 px$ resolution. NASNetLarge is the second largest model considered in the present study with 88,949,818 trainable parameters, compared to 138,357,544 trainable parameters in VGG-16. It is also the model that offers the best accuracy and F1-Score among all models considered in this study.

We find NASNetLarge to be robust against a wide range of ill exposure levels, as well as lossy compression, gross Gaussian and Poisson Noise. Nevertheless, it also suffers with S&P and Speckle Noise. As observe in Inception-ResNet-v2 [53], NASNetLarge also presents results that are not symmetric for under-exposure and over-exposure. Both on γ power transformed images and quantile-wise truncated samples, we notice that under-exposed images degrade the predictors' performance more than over-exposed images. Poisson Noise, mild miss-exposure, and Gaussian Noise show little impact.

NASNetMobile [71], which constitutes a smaller version (in terms of trainable parameters) of NASNet, takes

RGB images up to $224 \times 224 px$ as input. In comparison to its larger version, this model predictably presents inferior results under significant miss-exposure conditions, S&P and Speckle Noise, whilst mild under-exposure and overexposure, as well as Poisson Noise, apparently have a small effect on the image recognition accuracy and precision.

Overall, we observe that the image recognition models that obtained the highest accuracy in the original set of images, also obtain the best results when applied on the images that were distorted and manipulated. In a per distortion analysis, we notice that the ranking of the best performing models rarely switch positions. Exceptions to this condition are limited to extreme miss-exposure obtained through Gamma power transformation with $\gamma = [\frac{1}{8}; 8]$, Salt and Pepper Noise, and Speckle Noise. Nevertheless, under these conditions the accuracy obtained by some of the models renders them useless for practical applications.

The impact of the distortions are, in general, associated with the number of trainable weights in the network. NASNetLarge, as shown in Fig. 2, performs better than every other model evaluated in this study, regardless of the distortion applied to the input image. Inception-ResNet-v2 and Xception are slightly less robust. Those were the three models to show the best performance among the models presented in this paper.

Therefore, larger models hold better against distortions. The number of weights' importance is evidenced when taking into consideration the condition observed with NASNetMobile and NASNetLarge, which have identical base cells (architecture), but a great difference when it comes to the number of weights - although VGG-16, the largest model taken into account, provides us with a counter-example. Released in 2014, VGG-16 displays the lowest robustness of all the models considered in this study. That is to show that improvements such as residual blocks and separable convolutions combined with 1×1 2D convolutions, also granted a significant improvement to the classification models.

4.2 Noise Levels Versus Accuracy

We investigate the relationship between accuracy and noise intensity, taking into account the findings obtained in the first set of experiments. For this evaluation we explore the types of noise that have show strong decay all object recognition models. Due to its age and wide adoption on other visual perception, as well as being one of the simplest models (in terms of techniques and performance improvement tricks) we have chosen to work with VGG-16 [48].

Salt & Pepper and Speckle Noise have shown critical impact in the object recognition accuracy for all CNN models considered. Both are linked to sensor issues, amplifier issues, and, under certain conditions, to the physics of the amplifier itself. Salt & Pepper originates from defects on the sensor array which makes pixels become permanently on or permanently off. Speckle Noise is granular interference that is inherent to many image acquisition systems, such as the ones used in active radar, synthetic aperture radar (SAR), medical ultrasound and optical tomography. Speckle Noise is especially critical when the signal is sensitive to

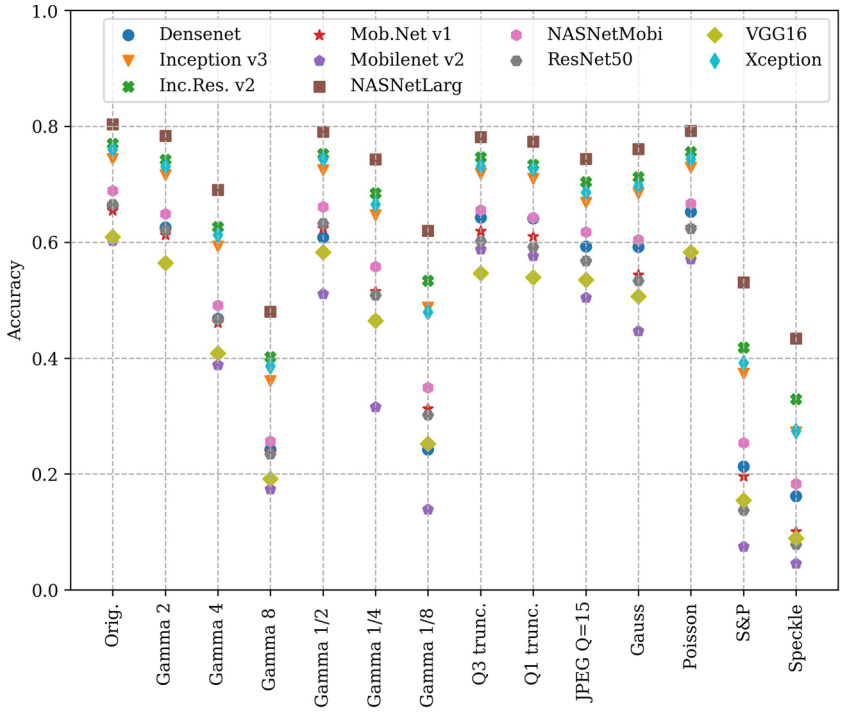

Fig. 2 Comparative Top-1 accuracy between image recognition models under evaluation

Table 3 Impact of noise severity on the Top-1 Accuracy

	Distortion level	Top-1	Top-5	F1-Score
Original	–	0.612	0.838	0.609
Salt & Pepper	0.05	0.436	0.671	0.439
	0.10	0.332	0.553	0.340
	0.15	0.263	0.464	0.275
	0.20	0.212	0.393	0.225
	0.25	0.173	0.336	0.186
Speckle	0.2	0.497	0.732	0.495
	0.4	0.307	0.519	0.315
	0.6	0.181	0.348	0.192
	0.8	0.115	0.243	0.125
	1.0	0.081	0.181	0.088

small variations, affecting the image representation in both spacial and spectral domains.

Table 3 shows the impacts of different levels of image distortion on the VGG-16 [48] model. We follow the same coloring scheme throughout the whole paper considering the impact over the original image set, that is, low impact ($\leq$ 10%), moderate impact ($>$ 10 and $\leq$ 30%), and critical impact ($>$ 30%).

For Salt & Pepper the amount of noise is defined as the proportion of pixels that are affected by noise. The amount of noise starts from 5% and goes up to 25%. We notice that at 5% the impacts on Top-1 Accuracy, Top-5 Accuracy and F1-Score are moderate. Anything higher than that results in critical impact considering that the prediction quality reduces to levels where an automated system would be unable to make reliable decisions. As the amount of noise increases we observe a further drop in the object recognition task which finally leads to levels were the object recognition system fails more often than it succeeds. The F1-Score provides a clear indication of both poor precision and poor recall.

For Speckle Noise, the amount of noise is controlled by the variance (S^2). While $S^2 = 1$ is a value commonly used to simulate Speckle Noise, we have tested with smaller S^2 values, in a strategy that is similar to the one used by Ren et al. [43]. Five sets of images with different speckle intensities were generated. The corresponding variances are 0.2, 0.4, 0.6, 0.8, and 1.0. A variance $S^2 = 0.2$ results in a moderate impact on the object recognition accuracy. Anything beyond this level of noise results in critical impact, dropping the prediction quality by more 30% in relation to the original non-distorted image set.

The results shown in Table 3 indicate that both Salt & Pepper and Speckle Noise are critical for robot vision. Robots and autonomous systems that rely on visual data to interact with the environment could become unreliable even under small amounts of noise. The development of robust algorithms requires further investigation on the mitigation of the impacts, CNN training strategies, as well as redundancy systems.

4.3 An Alternative Vision Pipeline

In order to recover miss-exposed images, we implemented the image restore model ReExposeNet [50] into the pipeline for image recognition. This model is intended to estimate the radiance of an inappropriately exposed image, a process that involves restoring and enhancing non-clipped pixels to improve visibility and color accuracy, as well as restoration strategies for regions where the signal has been clipped. It can synthesize large clipped parts in high-resolution images. ReExposeNet is a one-size-fits-all approach that can be continuously extended to a wide variety of levels of image miss-exposure. We used the model as released by its authors, without further fine-tuning.

To mitigate the noise damage we used the DnCNN-3 [68] denoising model. DnCNN-3 relies on residual learning and batch normalization to speed up the training process as well as boost the denoising performance. Zhang et al. claims to provide a single DnCNN model to tackle several general image denoising tasks, such as blind Gaussian denoising, single image super-resolution, and JPEG image deblocking. This model can be efficiently implemented to use GPU computing, which makes it suitable for real-time applications.

We chose to explore how the restoration pipeline impacts the VGG-16 [48] model. As shown in Section 4.1 VGG-16 [48] is highly susceptible to image distortion. Except for Poisson Noise, all image distortions resulted in an accuracy drop larger than 10% for this model. Table 4 compares the

Table 4 Top-1 Accuracy for VGG on distorted and restored images

Classification Model	[48]
Original Images	0.612
Gamma 8	0.175
Gamma 8 Restored	0.429
Gamma 1/8	0.236
Gamma 1/8 Restored	0.618
Gauss	0.508
Gauss Restored	0.497
Poisson	0.586
Poisson Restored	0.445
S&P	0.143
S&P Restored	0.141
Speckle	0.081
Speckle Restored	0.091

impact of damaged images and the effects of restoration on the VGG-16 [48] model.

The restoration pipeline provides an expressive gain under extreme miss-exposure. For extreme under-exposure, simulated via Gamma Power Transformation with $\gamma = 8$, the model pipeline approach is able to improve the Top-1 Accuracy from 0.175 up to 0.429. For extreme under-exposure, simulated via Gamma Power Transformation with $\gamma = \frac{1}{8}$, the pipeline approach was able to restore the object recognition accuracy from 0.236 up to 0.618. It is interesting to notice that this 0.618 Top-1 Accuracy is higher than the 0.612 accuracy obtained on the original image set.

For noisy images we notice that the inclusion of the denoising model actually worsened the results on both AWGN, Poisson Noise, and Salt & Pepper. For Speckle Noise, the restoration offered a small improvement. The differences between the Top-1 Accuracy obtained by VGG-16 for the noisy images and their restored counterparts are marginal. Although the restored images look better to the human eye, the object recognition models are unable to benefit from this visual improvement.

From these initial tests with a pipeline approach we can tell that, on one hand, the miss-exposure problem can be minimized by the use of image enhancement methods. The exposure enhancement model provides an expressive gain in conditions where pixels are clipped by both saturation and under-exposure. On the other hand, the image noise problem poses a more challenging task, where the denoising methods were unable to lead the object recognition model to better accuracy levels. Nevertheless, further investigation using a broad set of state-of-the-art denoising algorithms is needed to provide definitive evidence.

4.4 Experiments on Instance Segmentation

Other applications of perception for robotics, autonomous systems and machine perception may present similar degradation in accuracy when subjected to image degradation from noise or inadequate exposure. Object picking, localization and mapping, navigation, loop closure, obstacle avoidance, harvesting robots, and human-robot interaction systems often share the same basic building blocks with deep learning based image classifiers [7, 15, 24, 34, 35, 41, 56, 69]. In order to further understand how these common image distortions may impact autonomous and robotic systems, we explored how these affect instance segmentation.

In the instance segmentation task, the goal is to distinguish each instance of each object within the image at the pixel level. We evaluate how Mask-RCNN [16], a model designed to efficiently detect objects in an image while simultaneously generating a segmentation mask for each instance, performs under non-ideal conditions.

The visual results for instance segmentation of an urban scene using the Mask-RCNN model are shown in Fig. 3. This image shows a high contrast scene which represents practical and plausible situation in autonomous outdoor navigation. Overall, we notice a significant impact on the outcomes, especially under severe miss-exposure conditions and noise. We are able to identify the occurrence of both false-positives and false-negatives. Considering an hypothetical autonomous driving system, FPs and FNs, such as the ones shown in by this sample, may result in malfunctioning, insufficient data to take actions, or even in autonomous decisions that put lives in risk.

In order appearance Fig. 3 shows a properly exposed image of the scene (Fig. 3a); the instance segmentation results on the original image (Fig. 3b), JPEG over-compressed image (Fig. 3c), image affected by AWGN (Fig. 3d), image affected by Salt & Pepper Noise (Fig. 3e), image affected by Speckle Noise (Fig. 3f), image affected by Poisson Noise (Fig. 3g), image affected by Gamma $\frac{1}{2}$ (Fig. 3h), image affected by Gamma $\frac{1}{4}$ (Fig. 3i), image affected by Gamma $\frac{1}{8}$ (Fig. 3j), image affected by Gamma 2 (Fig. 3k), image affected by Gamma 4 (Fig. 3l), image affected by Gamma 8 distortion (Fig. 3m), image affected by Truncation Q_1 (Fig. 3n), image affected by and Truncation Q_3 (o). Each color represents one class label, defined as follows: lime represents a person; light blue represents a bicycle; gray represents a chair; yellow represents a potted plant; rose represents a vase; hot pink represents a umbrella; purple represents backpack; seafoam green represents a car; and white represents skis.

Figure 3b shows that, on the original image, Mask-RCNN is able to properly identify people, bicycles and traffic signs. In this condition, the autonomous system could rely on the segmentation results to perform localization, mapping and obstacle avoidance. JPEG over-compression, shown in Fig. 3c, shows no impact on the predictions. Additive White Gaussian Noise , shown in Fig. 3d, renders the application useless. The same holds for Fig. 3l and m which show underexposure generated by Gamma Power transformation with $\gamma = [4; 8]$. In Fig. 3k, which shows underexposure generated by $\gamma = 2$, we see that the amount of objects detected is significantly reduced.

Most of the image distortions result in an expressive increase in false-negatives. Going further, in Fig. 3f, i, and j, we observe that Mask-RCNN results in false-positives, including instances of objects like chair, backpack, vase, potted plants and skis. The severity of the impacts on robots and autonomous applications that rely on these systems is certainly up for discussion.

Many recent advancements in robotics rely on visual perception of the environment. Robotics and automation, human-robot interaction, human-machine interfaces and in-

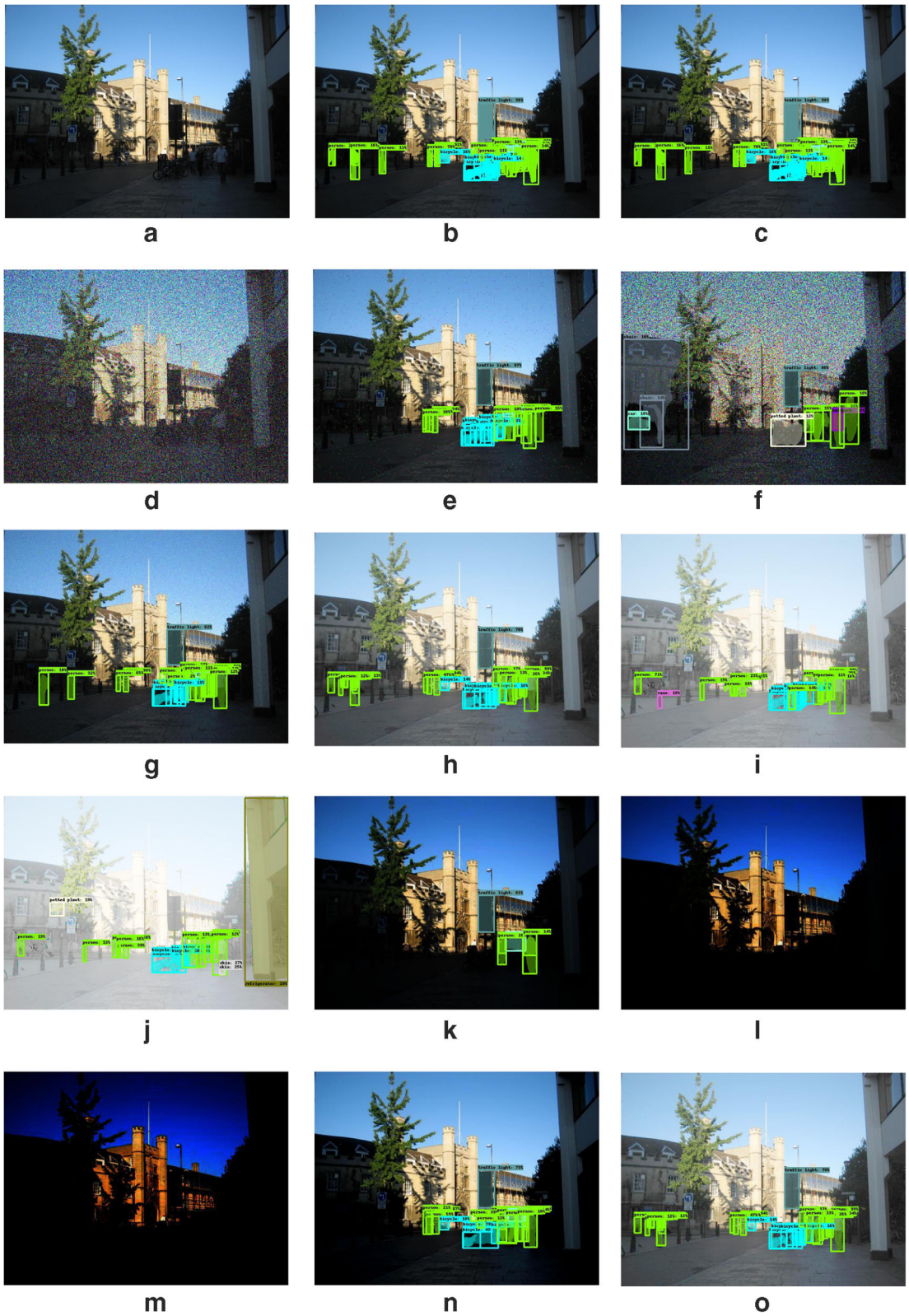

Fig. 3 Impact of the image distortions on a segmentation task: **a** Input Image, **b** Segmentation on the original image, **c** JPEG over-compression, **d** Additive white Gaussian Noise, **e** Salt & Pepper Noise, **f** Speckle Noise, **g** Poisson Noise, **h** Gamma $\frac{1}{2}$, **i** Gamma $\frac{1}{4}$, **j** Gamma $\frac{1}{8}$, **k** Gamma 2, **l** Gamma 4, **m** Gamma 8, **n** Truncation Q_1, and **o** Truncation Q_3. Class labels are as follows: lime - person; light blue - bicycle; gray - chair; yellow - potted plant; white - potted plant; rose - vase; hot pink - umbrella; purple - clothes; seafoam green - car; and white - skis

teraction, social and service robotics, medical robotics, unmanned systems, autonomous systems, cyber physical systems, and other related fields have benefited from the advancements in machine vision that are provided by deep neural networks. The findings in this experiment show that commonplace distortions may lead these systems to become unreliable or even hazardous. We therefore believe it offers a unique perspective in building high-confidence systems and emphasizes the importance of redundancy, as well as multi-modality sensors, including sound based sensors, depth sensors, and active sensors. Furthermore, we also believe that image pre-processing techniques and better image sensors also play a significant role for image based perception and could be applied to make these systems more reliable.

5 Discussion

First, the results obtained in the noise impact assessment raise a few questions regarding the practical applications of the image recognition models for vision-based tasks. All models are critically impacted by, at least, one type of distortion. While most object recognition models are little affected by mild mis-exposure or pixel truncation, harsh mis-exposure resulted in significant performance deprecation.

Gaussian Noise and Poisson Noise, as well as JPEG compression artifacts, have shown little impact on most object recognition models or on the segmentation model. Salt & Pepper and Speckle Noise, otherwise, have shown to be critical, even in small amounts. This behavior shows that further developments have to be done to improve the robustness of robot vision systems, especially considering that these are commonly occurring noise types that originate from sensor defects and faulty electronics. The current state-of-the-art in visual perception lacks robustness towards these distortions. Furthermore, we find that even small amounts of Salt & Pepper and Speckle Noise result in critical impact.

Corroborating common belief we observe that, when it comes to the reliability of deep-learning based models for vision tasks under non-ideal conditions, larger and newer models perform better. The inability to generalize to new sets of data has been explored before by Rech et al. [42], whom theorized on the potential causes for accuracy drops, and Fawsi et al. [10] whom performed an analysis of classifiers' robustness towards adversarial perturbations. Nonetheless, our results come to show that even simple and common image distortions are potential sources of error.

We have also explored an alternative robot vision pipeline, designed to mitigate mis-exposure and noise. We propose a modular approach that can be included in any robot vision pipeline that relies on visual perception using the visible spectrum of light. The initial results on the pipelined approach for under-exposed and over-exposed images are encouraging, proven that the restoration stage provides an expressive gain in the prediction accuracy. Otherwise, for noisy images, the proposed pipeline approach was able to offer only marginal gains. Nevertheless, we believe this approach could receive more attention in future work to explore alternative restoration models or even design restoration models that are more adequate for CNN based vision.

Finally, we show that same distortions that impact object recognition also impact other robot vision tasks. This was expected, considering that most object recognition is the most mature vision task at the present time. State-of-the-art segmentation models such as [3, 8, 20, 66, 70] mostly reuse components, building blocks, network architectures and model adjustment strategies that have already shown successful in other tasks.

6 Conclusion

This paper is associated with the assessment of CNN models for image recognition in real visual tasks. We propose a methodology to generate distorted images which replicate the effects observed in real applications and apply a wide set of metrics for classification. A comprehensive set of experiments was run in order to estimate the steadiness of modern Deep-Learning image recognition models when confronted with some of the most usual image distortions. The analysis was made taking into account a vast set of classifiers that had outstanding accuracy within the ImageNet Large Scale Visual Recognition Competition (ILSVRC), investigating the effects brought by poor exposure and over-compression, both usually connected to a low bandwidth and communication bottlenecks. The results of signal-dependent noise - often connected to film-grain and Speckle Noise - and signal-independent noise - often result from the physics of amplifiers and flaws on the sensor - are also examined.

The procedure presented in this paper is clear, simple, and reproducible. It can be used as a framework to appraise the performance of any image-based model in the fields of robotics and automation. The data is provided by a combination of classifier metrics, which evaluate the sensitivity, specificity, and robustness of the classification models according to the image distortions applied to the input. These offer a reliable representation of the model's capacity in real applications when used in conjunction with one another. The code to reproduce the results shown here is available at https://git.io/JUgIz.

Many relevant research topics arise from the assessment framework and the experimental results presented in this paper. In future work, one particularly fruitful avenue

for more investigation is to target task-aware adversarial perturbations to assess the robustness of deep neural networks more deeply. We believe this provides a relevant contribution both to academy and industry, especially when considering the growing popularity of technology that is based on visual scene understanding and labeling. Moreover, it would be relevant to analyze which features are being learned by the different models and how the distortions are affecting each one of them. This may indicate which models can be used for specific applications or which ones can be combined in an ensemble learning approach in order to increase the final accuracy.

We show that current image-recognition CNN models are significantly impacted by gross misexposure, such as the ones obtained by $\gamma = [\frac{1}{4}, \frac{1}{8}, 4, 8]$. In general, recent and large models are less affected. We find that the majority of the recognition models are little affected by mild misexposure or truncated pixel values. The same holds for Poisson and Additive White Gaussian Noise also show minimal impact on the accuracy of most evaluated models. Over-compression using the JPEG algorithm does not seem to impact the recognition models. Otherwise, all models have shown to be vulnerable to distortions originated from signal independent noise such as Speckle and Salt & Pepper.

As discussed before, vision is at the core of many recent advancements in robotics and autonomous systems. While the severity of false-positives and false-negatives in these systems depends on the role the perception plays in each task, it is undeniably relevant. The results we have obtained while applying our evaluation framework on several state-of-the-art classifiers show that commonplace image distortions may lead these systems to become unreliable, which in turn may lead to malfunctioning systems, premature wear of, ill controlled systems and even risk of accidents. We therefore believe this work offers a unique perspective in building high-confidence systems and emphasizes the importance of redundancy, as well as multi-modality sensors, including sound based sensors, depth sensors, and active sensors. Furthermore, we also believe that image pre-processing techniques and better image sensors also play a significant role for image based perception and could be applied to make these systems more reliable.

Acknowledgements We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan X Pascal GPU used for this research.

Author Contributions Cristiano Rafael Steffens, Lucas Ricardo Vieira Messias, Paulo Lilles Jorge Drews-Jr, and Silvia Silva da Costa Botelho all contributed equally to the work.

Funding This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

Availability of Data and Material The ImageNet dataset is available at http://www.image-net.org. Trained models for Mask-RCNN are available at https://github.com/matterport/Mask_RCNN. Trained models for object recognition are available at https://github.com/keras-team/keras.

Code Availability The code to reproduce the contributions of this paper is available at GitHub https://git.io/JUgIz.

Declarations

Consent to Publish This research involved no human subjects.

Consent to Participate This research required no study-specific approval by the appropriate ethics committee as it does not involve humans and/or animals.

Conflict of Interests The authors declare that they have no conflict of interest.

References

1. Abdelhamed, A., Brubaker, M.A., Brown, M.S.: Noise flow: Noise modeling with conditional normalizing flows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2019)
2. Chen, C., Seff, A., Kornhauser, A., Xiao, J.: Deepdriving: Learning affordance for direct perception in autonomous driving. In: The IEEE International Conference on Computer Vision (ICCV) (2015)
3. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European Conference On Computer Vision (ECCV), pp. 801–818 (2018)
4. Chollet, F.: Xception: Deep learning with depthwise separable convolutions. In: Proceedings of the IEEE Conference On Computer Vision And Pattern Recognition, pp. 1251–1258 (2017)
5. Diane, S.A., Lesiv, E.A., Pesheva, I.A., Neschetnaya, A.Y.: Multi-Aspect Environment Mapping with a Group of Mobile Robots. In: 2019 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConrus), pp. 478–482. IEEE (2019)
6. Drews, P. Jr., Hernández, E., Elfes, A., Nascimento, E.R., Campos, M.: Real-time monocular obstacle avoidance using underwater dark channel prior. In: 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 4672–4677 (2016)
7. Du, D., Qi, Y., Yu, H., Yang, Y., Duan, K., Li, G., Zhang, W., Huang, Q., Tian, Q.: The unmanned aerial vehicle benchmark: Object detection and tracking. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 370–386 (2018)
8. Emara, T., Abd El Munim, H.E., Abbas, H.M.: Liteseg: A Novel Lightweight Convnet for Semantic Segmentation. In: 2019 Digital Image Computing: Techniques and Applications (DICTA), pp. 1–7. IEEE (2019)
9. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. Int J Comput Vis **88**(2), 303–338 (2010)
10. Fawzi, A., Fawzi, O., Frossard, P.: Analysis of classifiers' robustness to adversarial perturbations. Machine Learn. **107**(3), 481–508 (2018). https://doi.org/10.1007/s10994-017-5663-3
11. Gao, F., Wang, C., Li, L., Zhang, D.: Altitude information acquisition of uav based on monocular vision and mems. J. Int. Robot. Syst. 1–12 (2019)

12. Gaya, J.O., Gonçalves, L.T., Duarte, A.C., Zanchetta, B., Drews, P. Jr., Botelho, S.S.C.: Vision-based obstacle avoidance using deep learning. In: 2016 XIII Latin American Robotics Symposium and IV Brazilian Robotics Symposium (LARS/SBR), pp. 7–12 (2016)
13. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the Kitti vision benchmark suite. In: Conference on Computer Vision and Pattern Recognition (CVPR), p. 8 (2012)
14. Gu, C., Sun, C., Ross, D.A., Vondrick, C., Pantofaru, C., Li, Y., Vijayanarasimhan, S., Toderici, G., Ricco, S., Sukthankar, R., et al: Ava: A video dataset of spatio-temporally localized atomic visual actions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6047–6056 (2018)
15. Ha, I., Kim, H., Park, S., Kim, H.: Image retrieval using bim and features from pretrained vgg network for indoor localization. Build. Environ. **140**, 23–31 (2018)
16. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2961–2969 (2017)
17. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
18. He, K., Zhang, X., Ren, S., Sun, J.: Identity mappings in deep residual networks. In: European Conference on Computer Vision, pp. 630–645. Springer (2016)
19. Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H.: Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv:1704.04861 (2017)
20. Hu, Y.T., Huang, J.B., Schwing, A.: Maskrnn: Instance level video object segmentation. In: Advances in Neural Information Processing Systems, pp. 325–334 (2017)
21. Huang, G., Liu, Z., van der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
22. Iocchi, L., Holz, D., Ruiz-del Solar, J., Sugiura, K., Van Der Zant, T.: Robocup@ home: Analysis and results of evolving competitions for domestic and service robots. Artif. Intell. **229**, 258–281 (2015)
23. Ito, K., Okano, T., Aoki, T.: Recent advances in biometrie security: A case study of liveness detection in face recognition. In: 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), pp. 220–227. IEEE (2017)
24. Jia, W., Tian, Y., Luo, R., Zhang, Z., Lian, J., Zheng, Y.: Detection and segmentation of overlapped fruits based on optimized mask r-cnn application in apple harvesting robot. Comput. Electron. Agric. **172**, 105380 (2020)
25. Karim, R., Islam, M.A., Mohammed, N., Bruce, N.D.: On the robustness of deep learning models to universal adversarial attack. In: 2018 15th Conference on Computer and Robot Vision (CRV), pp. 55–62. IEEE (2018)
26. Kohli, P., Chadha, A.: Enabling pedestrian safety using computer vision techniques: A case study of the 2018 Uber Inc. Self-driving car crash. In: Future of Information and Communication Conference, pp. 261–279. Springer (2019)
27. Kokil, P., Pratap, T.: Additive white gaussian noise level estimation for natural images using linear scale-space features. Circ Syst Signal Process 1–22 (2020)
28. Krizhevsky, A., Nair, V., Hinton, G.: The cifar-10 dataset. online: http://www.cs.toronto.edu/kriz/cifar.html **55** (2014)
29. Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Duerig, T., et al.: The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. arXiv:1811.00982 (2018)
30. Li, A., Thotakuri, M., Ross, D.A., Carreira, J., Vostrikov, A., Zisserman, A.: The ava-kinetics localized human actions video dataset. arXiv:2005.00214 (2020)
31. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference On Computer Vision And Pattern Recognition, pp. 2117–2125 (2017)
32. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft Coco: Common objects in context. In: European Conference on Computer Vision, pp. 740–755. Springer (2014)
33. Liu, W., Hu, J., Wang, W.: A novel camera fusion method based on switching scheme and occlusion-aware object detection for real-time robotic grasping. J. Int. Robot. Syst. 1–18 (2020)
34. Liu, Y.P., Yang, C.H., Ling, H., Mabu, S., Kuremoto, T.: A visual system of citrus picking robot using convolutional neural networks. In: 2018 5Th International Conference on Systems and Informatics (ICSAI), pp. 344–349. IEEE (2018)
35. Ma, L.Y., Xie, W., Huang, H.B.: Convolutional neural network based obstacle detection for unmanned surface vehicle. Math. Biosci. Eng. MBE **17**(1), 845–861 (2019)
36. Maity, A., Pattanaik, A., Sagnika, S., Pani, S.: A comparative study on approaches to speckle noise reduction in images. In: 2015 International Conference on Computational Intelligence and Networks, pp. 148–155. IEEE (2015)
37. Molina, M., Frau, P., Maravall, D.: A collaborative approach for surface inspection using aerial robots and computer vision. Sensors **18**(3), 893 (2018)
38. Osherov, E., Lindenbaum, M.: Increasing Cnn robustness to occlusions by reducing filter support. In: The IEEE International Conference on Computer Vision (ICCV) (2017)
39. Piyathilaka, L., Kodagoda, S.: Human activity recognition for domestic robots. In: Field and Service Robotics, pp. 395–408. Springer (2015)
40. Qian, K., Jing, X., Duan, Y., Zhou, B., Fang, F., Xia, J., Ma, X.: Grasp pose detection with affordance-based task constraint learning in single-view point clouds. J. Int. Robot. Syst. (2020)
41. Qiu, K., Ai, Y., Tian, B., Wang, B., Cao, D.: Siamese-Resnet: Implementing loop closure detection based on siamese network. In: 2018 IEEE Intelligent Vehicles Symposium (IV), pp. 716–721. IEEE (2018)
42. Recht, B., Roelofs, R., Schmidt, L., Shankar, V.: Do imagenet classifiers generalize to imagenet? arXiv:1902.10811 (2019)
43. Ren, R., Guo, Z., Jia, Z., Yang, J., Kasabov, N.K., Li, C.: Speckle noise removal in image-based detection of refractive index changes in porous silicon microarrays. Scientif. Rep. **9**(1), 1–14 (2019)
44. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-fei, L.: ImageNet large scale visual recognition challenge. Int. J. Comput. Vision (IJCV) **115**(3), 211–252 (2015). https://doi.org/10.1007/s11263-015-0816-y
45. Sánchez-Ramírez, E.E., Rosales-Silva, A.J., Alfaro-Flores, R.A.: High-precision visual-tracking using the imm algorithm and discrete gpi observers (imm-dgpio). J. Intell. Robot. Syst. **99**(3), 815–835 (2020)
46. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: Mobilenetv2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4510–4520 (2018)

47. Sharadqh, J.A.A.A., Ayyoub, B., Alqadi, Z., Al-azzeh, J.: Experimental investigation of method used to remove salt and pepper noise from digital color image. Int. J. Res. Adv. Eng. Technol. **5**(1), 23–31 (2019)
48. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556 (2014)
49. Soares, L.B., Weis, Á.A., Rodrigues, R.N., Drews, P.L., Guterres, B., Botelho, S.S., Nelson Filho, D.: Seam tracking and welding bead geometry analysis for autonomous welding robot. In: 2017 Latin American Robotics Symposium (LARS) and 2017 Brazilian Symposium on Robotics (SBR), pp. 1–6. IEEE (2017)
50. Steffens, C.R., Huttner, V., Messias, L.R.V., Drews, P.L.J., Botelho, S.S.C., Guerra, R.S.: Cnn-based luminance and color correction for Ill-Exposed images. In: 2019 IEEE International Conference on Image Processing (ICIP), pp. 3252–3256 (2019). https://doi.org/10.1109/ICIP.2019.8803546
51. Steffens, C.R., Messias, L.R.V., Drews, P.L.J., da Costa Botelho, S.S.: Can exposure, noise and compression affect image recognition? an assessment of the impacts on state-of-the-art convnets. In: 2019 Latin American Robotics Symposium (LARS), 2019 Brazilian Symposium on Robotics (SBR) and 2019 Workshop on Robotics in Education (WRE), pp. 61–66. IEEE (2019)
52. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A.: Inception-V4, Inception-Resnet and the impact of residual connections on learning. In: AAAI, vol. 4, p. 12 (2017)
53. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2818–2826 (2016)
54. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R.: Intriguing properties of neural networks. arXiv:1312.6199 (2013)
55. Szeliski, R.: Computer Vision: Algorithms and Applications. Springer Science & Business Media, Berlin (2010)
56. Teso-Fz-Betoño, D., Zulueta, E., Sánchez-Chica, A., Fernandez-Gamiz, U., Saenz-Aguirre, A.: Semantic segmentation to develop an indoor navigation system for an autonomous mobile robot. Mathematics **8**(5), 855 (2020)
57. Verma, R., Ali, J.: A comparative study of various types of image noise and efficient noise removal techniques. Int. J. Adv. Res. Comput. Sci. Softw. Eng. 3(10) (2013)
58. Voigtlaender, P., Krause, M., Osep, A., Luiten, J., Sekar, B.B.G., Geiger, A., Leibe, B.: Mots: Multi-object tracking and segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7942–7951 (2019)
59. Vono, M., Dobigeon, N., Chainais, P.: Bayesian image restoration under poisson noise and log-concave prior. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (IEEE), pp. 1712–1716. IEEE (2019)
60. van der Walt, S., Schönberger, J.L., Nunez-Iglesias, J., Boulogne, F., Warner, J.D., Yager, N., Gouillart, E., Yu, T.: The scikit-image contributors: Scikit-image: Image processing in Python. PeerJ **2**, e453 (2014). https://doi.org/10.7717/peerj.453
61. Wang, P., Huang, X., Cheng, X., Zhou, D., Geng, Q., Yang, R.: The apolloscape open dataset for autonomous driving and its application. IEEE Trans. Pattern Anal. Machine Intell. (2019)
62. Weber, F., Rosa, G., Terra, F., Oldoni, A., Drew-Jr, P.: A low cost system to optimize pesticide application based on mobile technologies and computer vision. In: 2018 Latin American Robotic Symposium, 2018 Brazilian Symposium on Robotics (SBR) and 2018 Workshop on Robotics in Education (WRE), pp. 345–350 (2018)
63. Weis, Á.A., Mor, J.L., Soares, L.B., Steffens, C.R., Drews-Jr, P.L., de Faria, M.F., Evald, P.J., Azzolin, R.Z., Nelson Filho, D., Botelho, S.S.D.C.: Automated seam tracking system based on passive monocular vision for automated linear robotic welding process. In: 2017 IEEE 15th International Conference on Industrial Informatics (INDIN), pp. 305–310. IEEE (2017)
64. Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1492–1500 (2017)
65. Young, K.Y., Cheng, S.L., Ko, C.H., Tsou, H.W.: Development of a comfort-based motion guidance system for a robot walking helper. J. Intell. Robot. Syst. 1–10 (2020)
66. Zhang, H., Wu, C., Zhang, Z., Zhu, Y., Zhang, Z., Lin, H., Sun, Y., He, T., Mueller, J., Manmatha, R., et al.: Resnest: Split-attention networks. arXiv:2004.08955 (2020)
67. Zhang, J., Hirakawa, K.: Improved denoising via poisson mixture modeling of image sensor noise. IEEE Trans. Image Process. **26**(4), 1565–1578 (2017)
68. Zhang, K., Zuo, W., Chen, Y., Meng, D., Zhang, L.: Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising. IEEE Trans. Image Process. **26**(7), 3142–3155 (2017)
69. Zhang, W., Zhou, H., Sun, S., Wang, Z., Shi, J., Loy, C.C.: Robust multi-modality multi-object tracking. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2365–2374 (2019)
70. Zhang, Z., Zhang, X., Peng, C., Xue, X., Sun, J.: Exfuse: Enhancing feature fusion for semantic segmentation. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 269–284 (2018)
71. Zoph, B., Vasudevan, V., Shlens, J., Le, Q.V.: Learning transferable architectures for scalable image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8697–8710 (2018)

Cristiano Rafael Steffens is a M. Sc. in Computer Engineering, and Ph.D. candidate in Mathematical Modeling at Universidade Federal do Rio Grande (FURG). His interests are computer vision, autonomous systems, intelligent robotics, and image-based perception. He has developed several automation solutions that explore image processing and machine learning for industrial environments. Steffens is an enthusiast of open-source, open datasets, interpretable deep-learning models, and reproducible science.

Lucas Ricardo Vieira Messias is an undergraduate student in Automation Engineering at the Universidade Federal do Rio Grande (FURG), Brazil, working on the intersection of machine learning, computer vision, and robotics. His work includes automation applied to agriculture, entrepreneurship, and nowadays, he develops solutions for perception in intelligent robotics.

Paulo Jorge Lilles Drews-Jr is a D.Sc. and M. Sc. in Computer Science at Universidade Federal de Minas Gerais (UFMG), Brazil, under the supervision of Prof. Mario Campos. His main research interests are robotics, computer vision, image processing, pattern recognition, and machine learning. He was a researcher at the ISR Coimbra. He was also a visiting researcher in the ASL at QCAT-CSIRO, Australia. Currently, he is an Assistant Professor at Universidade Federal do Rio Grande.

Silvia Silva da Costa Botelho is an Electric Engineer and M. Sc. in Computer Science at Universidade Federal do Rio Grande do Sul (UFRGS). She has a Ph.D. in Robotics, Informatics, and Telecommunications at Centre National de la Recherche Scientifique. She is a Full Professor at Universidade Federal do Rio Grande (FURG) and Director of the Center of Computer Science and Director of the Intelligent Robotics and Automation Group (NAUTEC). Her research is mainly focused on Intelligent Robotic and Automation, applied to Oil and Gas industry and Education.