International Journal of Machine Learning and Cybernetics (2022) 13:1065–1077
https://doi.org/10.1007/s13042-021-01435-0

ORIGINAL ARTICLE

Really natural adversarial examples

Anibal Pedraza[1] · **Oscar Deniz**[1] · **Gloria Bueno**[1]

Received: 30 April 2020 / Accepted: 22 September 2021 / Published online: 5 October 2021

Abstract
The phenomenon of Adversarial Examples has become one of the most intriguing topics associated to deep learning. The so-called adversarial attacks have the ability to fool deep neural networks with inappreciable perturbations. While the effect is striking, it has been suggested that such carefully selected injected noise does not necessarily appear in real-world scenarios. In contrast to this, some authors have looked for ways to generate adversarial noise in physical scenarios (traffic signs, shirts, etc.), thus showing that attackers can indeed fool the networks. In this paper we go beyond that and show that adversarial examples also appear in the real-world without any attacker or maliciously selected noise involved. We show this by using images from tasks related to microscopy and also general object recognition with the well-known ImageNet dataset. A comparison between these natural and the artificially generated adversarial examples is performed using distance metrics and image quality metrics. We also show that the natural adversarial examples are in fact at a higher distance from the originals that in the case of artificially generated adversarial examples.

Keywords Natural adversarial · Adversarial examples · Trustworthy machine learning · Computer vision

1 Introduction

Adversarial examples are one the most clear examples of the current flaws in deep learning methods. They are images that are altered so that neural networks fail to classify them, even though they are identical for a human. The phenomenon is illustrated in Fig. 1.

Most of the methods related to this topic are based on crafting adversarial examples with artificial (calculated by algorithms) and carefully selected noise to make the networks fall into potential missclassifications [2]. Other works are aimed at increasing robustness to such adversarial "attacks", so they are called adversarial "defenses". Parallel to this, there is an intense discussion about the provenance of this phenomenon [1, 3], due to its importance in terms of security and reliability of systems that could govern future self-driving vehicles, access control or medical diagnosis systems.

It seems clear that further analysis on this phenomenon is important, since there are evident flaws in deep learning models that are not taken into consideration with current research of "traditional" attacks and defenses. It has deep impact on performance and security in systems that need to be reliable. For example, as [4] states, a classifier that performs worse on a real-world distribution than on a hypothetical L_p radius ball of artificial perturbations, may be useless or not desirable.

On the other hand, it is not clear if these customized perturbations are relevant to critical systems such as autonomous vehicles. In works like [5], it is claimed that adversarial examples are not much of a problem for such systems. In that work the authors contend that the continuous oscillation of the angle and viewing size of the objects mitigates their occurrence. However, even though this kind of noise does not manifest in real-world scenarios so frequently, it can be crafted so that it puts models at risk [6]. In this context, our work shows that this noise is present in the real world (and causes adversarial examples as well).

In the main line of research for adversarial examples, attack algorithms are developed to get information about the model. If they have access to the internal parameters, they are called white-box attacks. When they get information only about the label decision they are called black-box attacks. Independently of this, both approaches try to figure out how to perturb the original image to induce the

✉ Oscar Deniz
Oscar.Deniz@uclm.es

[1] VISILAB, ETSI Industriales, Avda. Camilo Jose Cela, 3, 13071 Ciudad Real, Spain

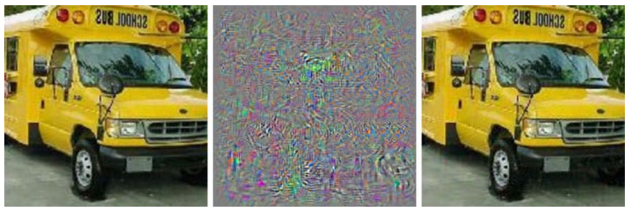

Fig. 1 Adversarial example. From left to right: original image (classified as "bus"), noise added, resulting adversarial image (classified as "ostrich") [1]

gradient of the model to flow towards another class different than the original. If the resulting class does not matter, it is called an untargeted attack. If a specified output class is enforced, it is called a targeted attack.

In this paper, we study the phenomenon of adversarial examples that are produced without the intervention of any algorithm. They could appear, for example, because of the noise present in an image that is naturally introduced when captured from real world using any camera-like device. As a way to differentiate them from (artificial) adversarial examples, these are called "natural adversarial examples". Therefore, this paper aims to develop further knowledge on natural adversarial examples. The main contributions are:

- Presentation of robust evidence of the existence of natural adversarial examples, including other domains different than those already studied in the state of the art .
- A new perspective to measure adversarial noise, using image quality metrics is proposed. This can be more informative than common metrics in the field (Euclidean distance L2 metric and Lp metrics in general).
- It is shown how image quality perturbations are observed in both natural and artificial adversarial examples, which can be useful to detect them.
- It is shown that natural adversarial examples introduce larger perturbations than those present in artificial attacks.

The rest of the paper is organized as follows. Section 2 contains a revision of the state of the art, regarding artificial and natural adversarial examples, as well as the use of image quality metrics to evaluate the noise introduced in adversarial images. Then, Sect. 3 details the datasets employed in this work, along with a revision of the quality metrics that are used. In Sect. 4, the experiments that are carried out are explained. The section also contains an extensive discussion about the results. Finally, Sect. 5 highlights the main conclusions that can be extracted.

Fig. 2 Samples from class 'scorpion' in the Imagenet-A dataset by [8]. Pretrained Imagenet models classify them as (from left to right, top to bottom): quill, snail, washing machine, spider web, manhole cover and nail

2 State of the art

The method proposed in [7] presented an approach to conceive what could be a first instance of natural adversarial examples. In their interpretation, an algorithm is employed to produce an artificial noise, but it is carefully selected to be produced only on relevant areas of the object. As a result, the adversarial examples are similar to the original images and artifacts in non-relevant areas of the image are avoided. In this case the adversarial examples are generated using the same techniques as most artificial examples, only with a restriction in the perturbed areas. In this sense, they should not strictly be called natural.

As far as the authors know, the first work that stated the term and focused on natural adversarial examples was [8]. In that work, the authors conceive the natural adversarial examples as images that are hard to classify by current state of the art models. In their method, the natural adversarial images are carefully (and manually) chosen to present evident features of a different class, so that errors are induced in this way. The authors developed a dataset inspired in ImageNet, called ImageNet-A, in which images from different sources (but same categories) were collected to produce adversarial results with unperturbed images. Some examples are shown in Fig. 2. They all should belong to "scorpion" class. However, the samples contain textures of other classes, so the model classifies them as other classes, such as quill, snail, manhole cover, spider web or washing machine. Although it is true that the images belong, in human-sense, to a different class, the fact is that they are very different from the original images which the network was trained with.

The work proposed in [8] introduces the concept of natural adversarial examples, but with a different perspective

than our work. There, the authors perform an exhaustive search of images from the same categories of ImageNet, but taken from other sources, such as from Flicker, iNaturalist and DuckDuckGo. The goal of this search was to discover images from one class that contained texture features of other classes, in order to deceive the model. In our work, we do not perform any "malicious" search of borderline examples but compare identical images of the same object. Some are predicted correctly but others fool the network. The only difference between these subsets is the noise patterns introduced by the capture device.

Regarding the artificial adversarial examples, some of the most successful methods are used in this work and detailed as follows. The DeepFool method was one of the first attacks that achieved good results in complex datasets [9]. This method is based on an estimation of the bounded hyperplane in which a classifier predicts the same class. Taking advantage from this information, the attack aims to exceed this border, so it is finally calculating a perturbation that produces an adversarial example. Carlini and Wagner (CW), is one of the most successful attacks in the field. Proposed in [10], it defines a cost function that calculates the perturbation to be introduced in the input of the network. The distance to the original sample has to be minimized so that the adversarial examples can be close to the original inputs, and, at the same time, fool the network decision. In consequence, they are difficult to detect visually or through defense techniques. Finally, HopSkipJump attack [11] (previously called Boundary Attack). Only by querying the model with slight perturbations, this algorithm is able to estimate the responses of the model to new perturbations and compute them efficiently.

As opposed to the generation of artificial adversarial examples (for which several algorithms such as HopSkipJump, Carlini and Wagner, DeepFool, Fast Gradient Sign Method, etc. are available), the study and generation of natural adversarial examples is in a more preliminary state. The seminal work [8] develops some procedures to guide the natural adversarial search in the input space domain. There, images collected from different sources were classified as natural adversarial examples depending on their prediction when compared with the groundtruth. However, the generation of this dataset was mostly hand-crafted and no fully automatic method has been proposed yet, as far as the authors know. In our work, a manual exploration of the microscopy field of view and the object video sources was needed to find which spots were suitable to generate the natural adversarial examples. However, the rest of the process including capture, analysis and metric quantification was automatic. Recently, some works such as [12] have proposed the use of image detectors as a previous step to select the areas that can potentially contain these background textures mentioned before.

Even non artificial, Euclidean transformations to the images can affect the network decision. This is shown in works like [13, 14], in which rotations and other smooth natural changes are applied. They proved that, without altering the contents or adding noise to the image, it is possible to produce adversarial examples with similar effects on the behaviour of the model. This is an interesting approach very close to out interpretation of natural adversarial examples. As no artificial noise is employed to produce the errors, these are more critical in real-time systems which are based on images, such as surveillance cameras.

All of the methods described above can not be considered as proper natural adversarial examples, since they induce the errors with iterative optimization algorithms. In our concept of natural adversarial examples, samples with real world noise (as the one from cameras, screens or even physical or natural changes in the objects and images) are better qualified to belong to this category.

In order to increase the neural network robustness against this phenomenon, some techniques have been developed. One example is training the model with other adversarial examples. This idea was first proposed in [3], aiming to increase the accuracy of the model by performing data augmentation with adversarial examples. After that, several works have expanded and applied this technique [15] for different datasets, models and attacks. However, the technique has already been circumvented by several works [16, 17], in which it is shown that any given adversarially trained model can be threatened again with adversarial examples crafted beyond the limits of those used for retraining.

In our work, the prospective benefit of its application would be even lower, since the natural adversarial cannot be restricted to any specific parameters (as opposed to artificial attacks), so the potential input perturbation space would make it very difficult to enhance the model behaviour against these noise patterns. Moreover, the applications of this technique can also have other negative side effects in model accuracy, such as inducing overfitting due to the large number of similar images that are employed in the training phase.

Another aspect that is under research is the ability to quantify the distortion that is introduced with adversarial examples. This is interesting in order to evaluate whether the noise is noticeable to the human eye or even by the machine. Different metrics to quantify the distance between adversarial examples and non-altered samples have been developed. Usually, traditional attacks and defenses employ L_p based distances (L_0, L_2 or L_{inf}) [4]. More recently, [18, 19] introduced an interesting approach in which full reference quality metrics are also valid (and relevant) to be used in this context. In [20], for example, a custom metric is developed to evaluate models in terms of patch adversarial robustness.

A more theoretical comparison among these types of metrics can be found in [21].

Some works have tested the effects of different noise sources and the accuracy for deep learning models. For example, [22] employs impulse noise to perturb images from Google's Cloud Vision API. This noise, which is not gradient-based (as usual in artificial examples algorithms), is able to fool the labels with similar capacity. This idea is more aligned with our work, since the kind of noise they use is more similar to random natural-occurring noise than in other works. However, with that method the resulting adversarial images are often noticeable to the human eye.

3 Dataset and methods

The focus of this work is to show the phenomenon of natural adversarial examples. For this reason we employ two datasets from very different fields. The first one is related to microscopy, in which precise object identification is needed. Then a more general purpose dataset such as ImageNet is employed.

In both cases, the main reasons we suggest that can raise the natural adversarial examples are:

- Noise: when an image is captured from the real world in a digitalization process, artifacts may be introduced.
- Small alterations in images: due to unstable results in deep learning classifiers, small changes in the perspective or angle of objects can be translated into missing detection or changes in the class decision [13, 14].

These datasets are employed to test different approaches of deep learning image identification, such as classification or region-based detection. To assess the difference between potential natural adversarial examples and legitimate images, state of the art image quality metrics, along with L_p based distances, are employed.

In the two scenarios considered in our work (microscopy images and object detection) objects and backgrounds are static. Besides, the camera angles, illumination and similar parameters also remain the same.

Considering the previous conditions in our experimental set up, we can reach the following conclusions. Depending on a specific noise pattern randomly captured from a camera, a deep learning classifier can predict an image incorrectly due to the strong reliance on low-level pixel patterns. This problem can also appear in a scenario in which a single photo of an object is taken. Depending on the noise pattern in this specific instance, the neural network can predict two identical objects correctly or not, independently of the accuracy of the model. This failure scenario without an evident cause is what we call natural adversarial examples.

In order to give a more detailed description of the natural adversarial examples, we now explain the taxonomies in which they can be enclosed. According to the categories proposed in [23], they belong to the visual recognition field, specifically applied to Convolutional Neural Networks. No support from the game theory was employed in this method. Regarding the kind of data, since it is employed with test images it should be classified as an evasion attack. Regarding the attacker's knowledge, it should be considered as black box, since these natural adversarial examples can be produced with the only knowledge of the model predictions, and no further knowledge about the architecture or internal parameters is needed.

Regarding the attack specificity, it is indiscriminate, also known as untargeted. The underlying reason for this is that it is difficult to force the output to a specific class without a fully automated method to control the prediction gradient. As a result of this, the method mostly affects to the false negatives, as samples from a considered class will be predicted as another instead. Conversely, if the attack was targeted, it would affect the false positive metric. Regarding the attack mode, since it has been designed as a standalone technique, it can be considered a non-colluding method. Finally, regarding the evaluation approach, it has been developed as an experimental method in which the results are discovered when testing a model.

3.1 Microscopy

We focus initially on a microscopy problem, in which the instances of the different classes are very similar in appearance. The images chosen represent diatoms, a kind of algae of interest in biology for water quality assessment [24]. Biologist take slides from riverlands to count and label these organisms. This process is usually performed manually. For this reason, machine learning techniques are being applied to get more accurate and faster identification. In Fig. 3, an example of these organisms is shown. The objective is to detect and classify the objects (diatoms), since there are myriads of different classes, depending on their shape and other representative features.

In order to process these slides, images are taken from the microscope in grayscale due to the nature of the staining process. An extensive experimentation about the topic, considering classic and deep learning techniques can be found in [25]. In our experiments, a YOLOv2 [26] object detector was trained. The source of data and parameter details were taken from the referenced work. The final model achieved 0.73 mAP, which can be considered a good detection rate because of the difficulty of this problem.

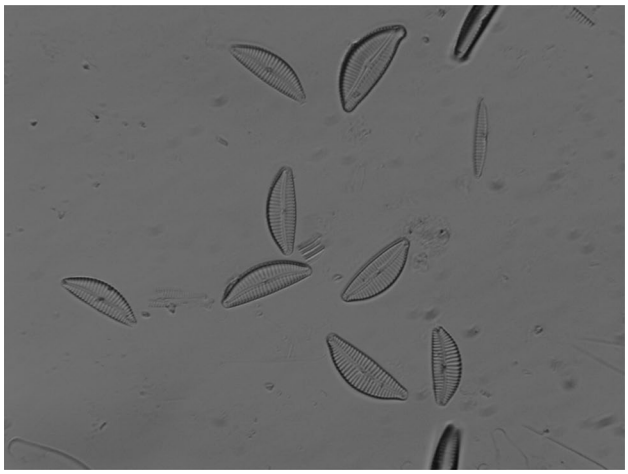

Fig. 3 Example of a diatom slide under the microscope

3.2 ImageNet

ImageNet [27] is one of the most used datasets in computer vision and deep learning. It contains 1000 classes of common objects, with more than a million of images to be trained with. For this reason, it has been a relevant dataset also for adversarial attack and defense evaluation, when applied with the most useful models (Inception [28], VGG [29], ResNet [30], YOLO [31]). Figure 4 shows a representation of some of the categories included.

In the case of our experimentation, a pretrained Xception architecture [32] is selected. It is one of the best and most efficient networks, achieving 0.79 Top-1 and 0.945 Top-5 accuracies.

3.3 Quality metrics

Image quality assessment is an important field to characterize with how well an image represents what is actually perceived from the original visual source. The field contains several techniques to provide a measure of fidelity similar to that of human vision [33].

When applied to adversarial research, distance metrics are provided to show a measure of the real difference between the original image and the adversarial. This allows to compare a method with others, since lower is better and, below a certain threshold, perturbations are no longer visible to human eye. To evaluate the difference between a reference image and natural adversarial examples, L_p based metrics are used, as summarized in Table 1. In our experiments, L_2 metric is used along with the rest of quality metrics exposed onwards.

As commented previously in the state of the art, image quality metrics have been recently proved recently to be useful to evaluate adversarial images, as such metrics are more aligned with the way a human perceives an image. For this reason, the following metrics are employed. They have been chosen as the best metrics to measure distance

Fig. 4 Composition with some of the Imagenet classes

Table 1 L_p norm metrics

Metric	Calculation	Explanation
L_0	Non-zero elements	Number of perturbed pixels
L_2	Euclidean distance	Distance in the image space
L_{inf}	Largest value	Highest perturbation at any pixel

between adversarial examples and original images according to human perception [18].

- Peak Signal to Noise Ratio (PSNR) [34]: this metric quantifies the maximum possible signal for the reference in comparison with the power of the distortion noise in the target image. It is measured in decibel (*dB*), since logarithmic expressions are used to calculate it. For example, regarding image quality degradation, PSNR values range from 30 to 50 *dB* in 8-bit representation (usual for each RGB channel in color images, or the whole image when grayscale). This metric is defined in Eqs. (1) and (2).

$$MSE = \frac{\sum_{M,N}[I_1(m,n) - I_2(m,n)]^2}{M * N} \tag{1}$$

For the Mean Squared Error (Eq. 1), the quadratic difference of two images I_1 and I_2 is calculated for each pixel coordinate m, n in the whole pixel space M, N.

$$PSNR = 10log_{10}\left(\frac{R^2}{MSE}\right) \tag{2}$$

In Eq. (2), R^2 represents the maximum fluctuation of the image space (e.g. 2^8 for a 8-bit unsigned codified image) and *MSE* represents the Mean Squared Error in Eq. (1).

- Structure Similarity Index Method (SSIM) [35]: this metric evaluates changes in the perception through the structural information, expressed as the relationship between pixels that are spatially close or interdependant in any way. For this purpose, it defines some concepts such as luminancy masking (to evaluate the distortions that affects the edges of the image) and contrast masking (to evaluate the distortions produced in the textures of the image). Finally, the metric is a pondering of all these criteria. This metric is defined in Eq. (3).

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \tag{3}$$

For a pair images x and y, μ_x and μ_y are the averages of x and y, σ_x^2 and $sigma_y^2$ are the covariances of x and y, σ_{xy} is the covariance of x and y. Finally, c_1 and c_2 are two variables two stabilize the division, when a weak denominator is present.

- Information Fidelity Criterion (IFC) [36]: this metric is based on the information shared between the reference and target images with a particular vision. It detects the statistical information contained in the target image about the reference. In contrast, most of the previous methods relied only on the information that could flow from to reference to the target in a shared channel.
- Visual Information Fidelity (VIF) [37]: it is a metric derived from the Information Fidelity Criterion (IFC) presented previously by the same authors. Both methods measure the amount of information shared between the reference and target images, but the former is more advanced. It does not only quantify the shared information but it also takes into account the rest of information that can be extracted from the target that is present on the reference.
- Most Apparent Distortion (MAD) [38]: this metric was developed to take into account different aspects proposed by other metrics, at the same time. It joins the appearance-based strategy (looking for changes in local statistics of spatial frequency components), as well as an estimation of the perceived distortions (using local luminance and contrast masking). It is important to notice that this metric was the best correlated with human perception for adversarial examples, according to [18].

4 Experiments and results

In this section, the two main experiments carried out in this work are described. For each one, the environment and conditions are detailed, along with a summary of the results and the visualization of the most interesting cases.

First, in both microscopy and object detection scenarios, a camera is placed at a fixed point. In the case of the microscopy, the camera is mounted in a specifically designed piece to attach it to the base. In the other case, the webcam has a base that allows to place it in a table and orient it in the desired direction. Then, a static object is placed in front of the camera, in order to take 100 frames of video. After that, images are classified with the corresponding neural network. Depending on the prediction output, they are categorized as regular images or natural adversarial examples. Using the subset of regular (non-adversarial) images, artificial adversarial examples are crafted (CW, HopSkip, ...). Finally, for each regular or adversarial subset, Euclidean and quality metrics are extracted and summarized.

4.1 Microscopy dataset

When evaluating an instance in microscopy, as performed with usual non-expert image labeling, the quality of the image plays an important role in successful identification.

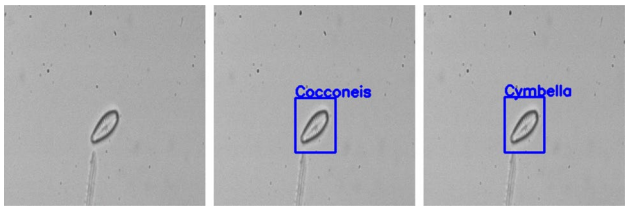

Fig. 5 Different classifications. The images are taken in consecutive frames

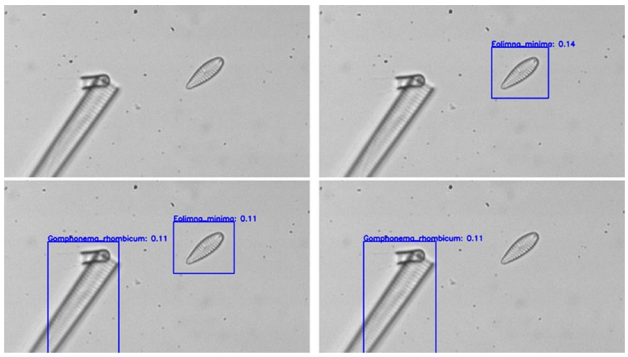

Fig. 6 Detection flickering. Instances are detected and missing in consecutive frames with imperceptible differences in background noise

However, in automatic identification, the task is conditioned by the following aspects:

– Classification of images with similar quality metrics, some correctly classified while others not. Considering that they are close in quality metrics and perceptually qualified by a human expert, the behaviour of the automatic classifier is not reliable. This is shown in Fig. 5. Although the three images have similar values in quality metric and are almost identical to human perception, the algorithm gives different results.
– Detection that flickers or disappears, as shown in Fig. 6. The difference between both images are only minor camera noise imperfections not visually appreciable. However, depending on the conditions, actual instances could not be detected due to this apparently random behaviour. This effect is also observed in a wider range of application fields, such as in pedestrian detection, as shown in Fig. 7, where missing detections are produced in consecutive frames.

The experiment conditions are set up carefully for reproducibility purposes. The equipment employed is a Brunel SP-30 optical microscope, with an Imaging Source DMK 72BUC02 camera. A biological slide is then placed, adjusting the augmentation and focus properly so that a single instance is clearly visible. Then, recording settings are defined to capture images at 1280×720, and the central region with size 416×416 is cropped to isolate the object.

Usually, when image classification tasks are performed, a single shot is taken so that this image will be used in the dataset, to train or test a model, after being labeled. However, in our experiment, a video is captured instead of a photo. Our aim is to observe how the noise from the camera acts as a natural perturbation, making the detector fail in consecutive frames. These natural adversarial examples expose this behaviour: the model misses the detection or confuses the correct class with another, even when the environment is static (fixed camera in the microscope, and static object in the slide). In this case, natural adversarial examples exhibit the same behaviour as artificial adversarial

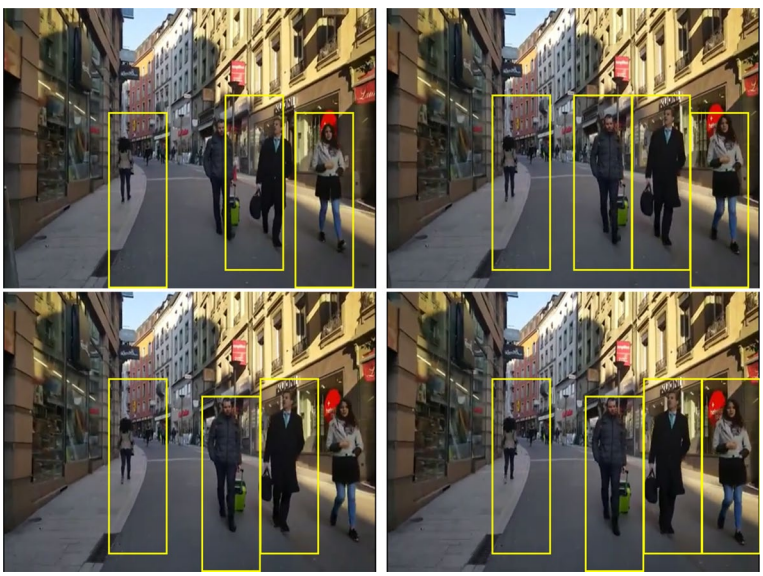

Fig. 7 Detection flickering. The pedestrians on the middle and right of the picture are not detected in these consecutive frames

examples: small and imperceptible noise and changes in the image affect the final decision of the classifier.

To evaluate this natural adversarial phenomenon, we take the first frame of the video as reference. Since all the frames are almost identical, they should give the same output when fed to the model. In our experimentation, we observe that this was not the case. Instead, the noise produced by the camera makes the model to switch between different class predictions, or even miss the object of interest.

Firstly, we isolate each of these frames, so we are able to measure the distance among them, using different metrics described in Sect. 3.3. This allowed us to quantify the difference between frames that are classified correctly and those that are not. The latter could be considered as "sucessful" natural adversarial. Out of 90 frames, 33 produced wrong classifications, confusing "Cymbella" class samples with "Cocconeis" objects, even when the images are identical. Therefore, the natural adversarial success rate was 63%.

Secondly, we take each frame and try to craft an artificial example. For this purpose, we use a recent and powerful black-box attack method called HopSkipJump [11]. This method only requires calls to the model with various inputs to guess the trend in the model gradient and, therefore, produce the adversarial examples with the specific noise for the chosen target. As a black box attack, it is adequate to be evaluated on different kinds of architectures, such as detectors, classifiers, segmentation models... In our experimentation, the attack is untargeted. This means that the algorithm tries to fool the network with any class, the one closest in the data manifold, so the gradient tends to a class with the smallest changes possible. When the attack is targeted, the algorithm is forced to fool the network in a specific class. This has some benefits but often leads to more distance between the original and the adversarial. Finally, for computation purposes, the maximum number of iterations allowed was 100, as a valid standard provided in the original publication. For the same 90 frames, the artificial attack correctly fools 69, so the attack rate is 77%.

The results of the experiment described above are shown in Table 2. In order to interpret the results, it is important to note that for SSIM, VIF, IFC greater values indicate more similarity, while for MAD, PSNR, L2 lower values indicate that the images are more similar. For the natural adversarial examples, the quality metrics show that the noise introduced by the camera is remarkable. The structural similarity is high, as the image throughout the frames is, essentially, the same. However, both VIF and ICF are indicating that the noise introduced by the camera is affecting the whole image. Regarding the L_2 distance, it is not so high (6.8) considering that other metrics suggest it could be larger.

In the artificial adversarial examples, the change is notorius. Both SSIM and VIF are close to 1, meaning that the images are nearly identical. Also, the Most Apparent Distortion is much lower, so the perturbations are minimal. Although PSNR values are similar, L_2 distance is also decreased, although the high standard deviation means that there were some examples that required minimum distortions to be fooled, while the hardest ones were close to the distance obtained for the natural adversarial examples. In conclusion, for these images and microscopical magnification, the noise required for the artificial method was smaller than the one naturally produced by the camera.

Table 2 Adversarial quality and distance metrics in microscopy

Metric	Natural	Artificial
SSIM	0.60 ± 0.001	0.97 ± 0.05
VIF	0.17 ± 0.002	0.90 ± 0.12
IFC	0.13 ± 0.004	0.52 ± 0.19
MAD	74.26 ± 2.20	4.41 ± 12.2
PSNR	44.39 ± 0.24	55.98 ± 6.37
L2	6.80 ± 0.02	2.07 ± 1.63

In order to interpret the results, it is important to note that for SSIM, VIF, IFC greater values indicate more similarity, while MAD, PSNR, L2 lower values point that the images are more similar as well

4.2 ImageNet

In the case of this dataset, a similar set-up is proposed: a fixed camera is placed looking at a single object. The main example in this case is a single pen, which is centered in the image. The selected device is a Logitech C525 HD webcam, capturing RGB color images at $640 \times 480 \times 3$ resolution, which is enough for the input required by the network ($299 \times 299 \times 3$ pixels).

First, the device is placed on a fixed and static position, so any variation is only due to noise pattern inherent to the camera (if any). Once the environment has been set up, the object is placed in the scene. After that, a live visualization of the classification is performed, studying how the object position affects the decision. This is performed until a proper place is found (according to the goal of this work). That is, the output starts to flicker or jump between different classes. Then, a hundred consecutive frames are captured and classified, to further develop the rest of the methodology.

After that, the quality metrics are calculated for a set of 100 frames extracted from the scene. Then, artificial adversarial examples are also generated. The conditions are the same than in the previous experiment. In this case HopSkipJump, Carlini and Wagner and DeepFool untargeted attacks are employed, limited to 100 iterations to craft the adversarial examples. This is enough to achieve a 100% adversarial success, so all the images are erroneously classified into other classes. The results from both kind of adversarial examples are shown in Tables 3, 4 and 5. These

three experiments are performed for different sets of images, whose groundtruth corresponds to categories ballpoint and whistle. The first and second one, contain a ballpoint with slightly different angles, which turns to different "natural adversarial examples" and class predictions. The third one substitutes the ballpoint object for a whistle, in a position which also conducts the model to adversarial predictions.

For the Imagenet dataset, the camera produced very small noise in the images, so they are practically identical. This is observed in the nearly 1 (0.96) value for SSIM and the very low value for L_2. Also, the MAD is very low in this case. This result is surprising, as the metrics indicate that all the frames are almost identical. However, as it will be shown in the rest of this section, the outputs given by the network are rather variable.

For the case of the artificial adversarial examples, the introduced distortions are even lower. The values provided by the quality metrics show that mostly imperceptible differences between the images are introduced. Very high values are present for SSIM, VIF and mostly in IFC. It is important to notice that they are in the range of what is normal for this kinds of attacks, which need very few pixels to induce the errors. Considering for example the L_2 distance, the authors of the HopSkipJump attack report similar values (around 2.3) when their algorithm is used in very deep convolutional networks for the ImageNet dataset. This is due to the complexity of the model (Xception, as many others tested on Imagenet, is a very deep architecture) and the large size of the images. In some cases, up to 20,000 iterations are needed to produce very low perturbations. For our purpose, that was not necessary, considering that the adversarial success rate was already 100% with 100 iterations.

Next, a comprehensive visualization and summary of the different situations encountered with this dataset is performed. The main classifications provided for consecutive frames are shown in Fig. 8. The frames are classified into several different classes, mostly as a "fountain pen", which is in the same synset (concept or family of concepts, used regularly in the Imagenet terminology) of similar classes that this object could be classified, such as "quill" or "ballpoint". If all the frames were classified as one of these three classes (considering a Top-5 classifications as usual in the state of the art), that could be considered a normal behaviour for this kind of networks pretrained on ImageNet. However, the fact is that revolver appears on 60% of the images as the Top-1 prediction (and rifle is also present in the Top-5 for most of them), which can be considered as a (natural) adversarial behaviour.

After applying the artificial attack, all the examples are successfully fooled to other classes, such as: "paddle", "projectile", "letter opener" and "speedboat". It is important to notice that they are the same classes achieved for natural adversarial examples in similar positions of the objects.

Fig. 8 Natural adversarial of a ballpoint which are classified as a fountain pen, but also as a revolver!

Again, as in the previous experiment, this suggests that these classes in particular are close in the gradient manifold for both natural and artificial adversarial examples. This is observed in Table 7, where all the classes predicted by the natural and artificial adversarial examples are shown. In order to have the same number of examples in each column, the average of a hundred executions has been calculated for the natural adversarial examples, with different subsets in each case. Moreover, the significance of each distribution is calculated with the Pearson coeficient as well as the p-value. The results, shown in Table 8, indicate that positive correlation is observed, with high significance for the Carlini and DeepFool attacks.

In order to check the artificial noise that is introduced by the artificial attacks, the difference between an original and a perturbed image is calculated. The same operation is performed with a natural adversarial. In this way, we compare both natural noise and adversarial noise in Fig. 9 for the ballpoint experiment and in Fig. 10 for the whistle experiment. As it can be observed, most of the distortions (lighter areas) are in the area where the object is present. This kind of noise follows the same pattern as the adversarial examples generated in [7]. Also, it is important to note that adversarial noise is larger (lighter) than natural noise, as extracted from metrics. Moreover, the perturbed images have also been extracted, in order to check if any of the perturbations are visually percetible. They are shown in Fig. 11 for the ballpoint experiment and in Fig. 12 for the whistle experiment.

It is important to notice that very similar images are provided in the Imagenet training set, so the examples obtained in this work can also match the concept of natural adversarial provided by [8]. In Figure 13 we show some of them.

Checking other positions and angles as explored in [13], it is possible to find other natural adversarial examples in which, with an static camera, the classification fluctuates among screw, paddle, projectile, speedboat or letter opener. In Fig. 14 some of these cases are shown. The objects for the whistle experiment show the same behaviour in Fig. 15 where some of them are presented. A whistle, with different perspectives, can be classified by the network as a "cup" (also a screw! sometimes), when the previous "identical"

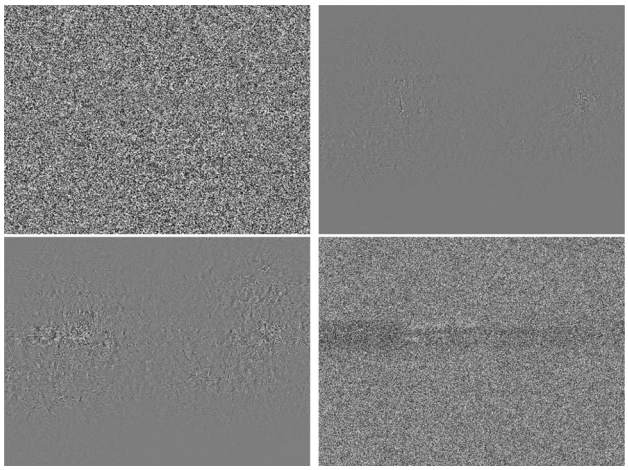

Fig. 9 Absolute difference between original and adversarial images for the ballpoint experiment. From left to right, top to bottom: Natural, CW, DeepFool, HopSkipJump

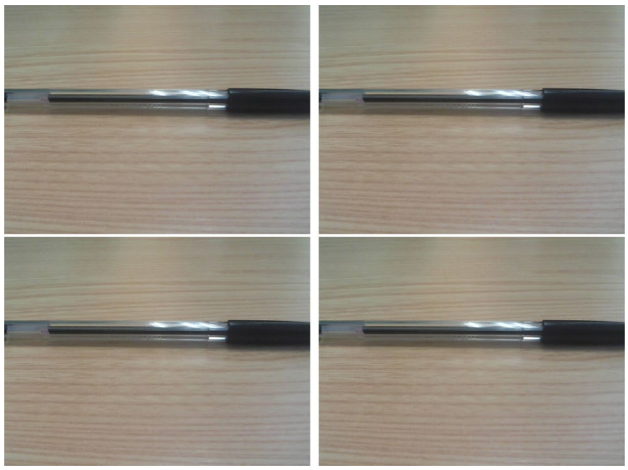

Fig. 11 Visualization of reconstructed images from adversarial attacks for ballpoint experiment. From left to right, top to bottom: Natural, CW, DeepFool, HopSkipJump

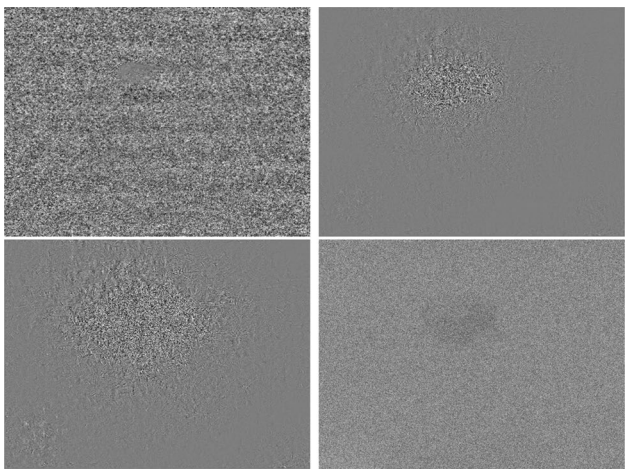

Fig. 10 Absolute difference between original and adversarial images for the whistle experiment. From left to right, top to bottom: Natural, CW, DeepFool, HopSkipJump

Fig. 12 Visualization of reconstructed images from adversarial attacks for whistle experiment. From left to right, top to bottom: Natural, CW, DeepFool, HopSkipJump

frame was classified as a "whistle". The angle in which the object is presented makes it look like a kind of cup. For this reason, this adversarial can be considered similar as in the line of work of [8], where objects with "features" of other objects are useful to craft natural adversarial examples.

In order to increase the significance of the results, an additional experiment has been carried out. There, a new camera has been employed (a Logitech C170 model), but the rest of environmental conditions remain the same. Moreover, we keep the experimental scheme applied in the rest of the work. That is: a hundred of frames from a single object (a ballpoint, for better comparison with the previous experiments) were taken, with fixed spatial position and no other source of illumination noise or any other kind. In this case, nearly a half of the images were predicted by

Fig. 13 Imagenet training samples similar to the object in our experiment

Fig. 14 Collection of natural adversarial examples in frames for a fixed position (speedboat, paddle, letter opener and projectile)

Fig. 15 Whistle in different angles being confused with a cup and a screw!

Table 3 Adversarial quality and distance metrics in ImageNet for Experiment 1

Metric	Natural	Hopskip	CW	DeepFool
SSIM	0.96	0.95	1.00	0.99
VIF	0.68	0.88	0.98	0.97
IFC	0.34	0.57	0.73	0.72
MAD	43.37	9.71	0.00	0.00
PSNR	31.88	38.09	47.75	47.69
L2	3.89	2.27	0.69	0.69

In order to interpret the results, it is important to note that for SSIM, VIF, IFC greater values indicate more similarity, while for MAD, PSNR, L2 lower values indicate that the images are more similar

the model (we also keep the pretrained Xception) with a wrong class, despite all the set of images being visually identical. The quantified results (Table 6) show that there are only slight differences between this experiment and the previous ones. Specifically, L2 distance and quality metrics remain in the same ranges for both the natural and artificially crafted adversarial examples. As a result, it is shown how the appearance of this kind of noise turns out

Table 4 Adversarial quality and distance metrics in ImageNet for Experiment 2

Metric	Natural	Hopskip	CW	DeepFool
SSIM	0.96	1.00	1.00	0.99
VIF	0.71	0.98	0.98	0.98
IFC	0.35	0.77	0.72	0.71
MAD	33.28	0.00	0.00	0.00
PSNR	36.24	49.02	49.63	49.63
L2	3.97	0.81	0.68	0.69

In order to interpret the results, it is important to note that for SSIM, VIF, IFC greater values indicate more similarity, while for MAD, PSNR, L2 lower values indicate that the images are more similar

Table 5 Adversarial quality and distance metrics in ImageNet for Experiment 3

Metric	Natural	Hopskip	CW	DeepFool
SSIM	0.95	1.00	1.00	0.99
VIF	0.68	0.98	0.98	0.98
IFC	0.35	0.83	0.79	0.78
MAD	19.71	0.32	0.00	0.00
PSNR	42.94	52.50	52.63	52.64
L2	2.65	0.80	0.67	0.68

In order to interpret the results, it is important to note that for SSIM, VIF, IFC greater values indicate more similarity, while for MAD, PSNR, L2 lower values indicate that the images are more similar

Table 6 Adversarial quality and distance metrics in ImageNet for a new camera set-up

Metric	Natural	Hopskip	CW	DeepFool
SSIM	0.92	1.00	1.00	0.99
VIF	0.47	0.99	0.98	0.98
IFC	0.18	0.86	0.79	0.78
MAD	43.86	0.00	0.00	0.00
PSNR	36.51	50.27	50.41	50.33
L2	3.09	0.71	0.70	0.71

In order to interpret the results, it is important to note that for SSIM, VIF, IFC greater values indicate more similarity, while for MAD, PSNR, L2 lower values indicate that the images are more similar

to be independent from the considered device or kind of images.

5 Conclusions

In this work we present "really natural adversarial examples" that may be a very powerful tool to continue exploiting deep neural networks flaws in high dimensionality problems such as ImageNet. It should be kept in mind that they are produced in an environment which should be

Table 7 Results for artificial and natural adversarial examples for experiments with Ground truth = ballpoint

Class	CW	DeepFool	HS	NAT
fountain_pen	1	0	10	0.46
letter_opener	2	0	2	0.39
paddle	13	14	3	20.99
projectile	4	3	3	3.94
puck	0	0	0	0.4
rubber_eraser	3	3	5	1.87
screw	19	19	17	13.01
screwdriver	0	0	4	0
speedboat	7	10	5	7.94

NAT is the mean of 100 executions to have the same number of samples as for artificial attacks

Table 8 Pearson correlation between artificial and natural adversarial examples

Method	CW	DeepFool	HS
Pearson	0.87	0.89	0.28
p-value	0.002	0.0013	0.47

considered static, since all the experiments are extracted from videos captured by fixed cameras for different sources (microscopy, real-world images).

From the results observed in Sect. 4.1, we first demonstrate the existence of the natural adversarial examples, as defined in this work. Later, they are extrapolated to a widely used dataset in computer vision, such as Imagenet. The first significant conclusion we can extract is that the noise introduced by the webcam for real world images was much lower than the introduced for microscopic camera (as supported by image quality metrics). Even in this case, a deep convolutional network (pretrained on the Imagenet dataset) could not perform accurately on these images.

For this reason, we suggest that further development for model defenses is necessary. Not only for spatial perturbations but for natural noise patterns. With models that are more robust to higher amounts of noise, they could be able to avoid the occurrence of these natural adversarial examples, being more stable in their behavior.

Acknowledgements This work was partially funded by the Spanish Ministry of Economy and Business with reference SBPLY/17/180501/000543 and by the Autonomous Government of Castilla-La Mancha; as well as the Postgraduate Grant FPU17/04758 from the Spanish Ministry of Science, Innovation, and Universities.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature.

References

1. Szegedy C Zaremba, W Sutskever, I Bruna, J Erhan, D Goodfellow, I Fergus R (2013) Intriguing properties of neural networks. arXiv preprint. arXiv:1312.6199
2. Serban AC, Poll E, Visser J (2018) Adversarial examples-a complete characterisation of the phenomenon. arXiv preprint. arXiv:1810.01185
3. Goodfellow IJ, Shlens J, Szegedy C (2014) Explaining and harnessing adversarial examples. arXiv preprint. arXiv:1412.6572
4. Gilmer J, Adams RP, Goodfellow I, Andersen D, Dahl GE (2018) Motivating the rules of the game for adversarial example research. arXiv preprint. arXiv:1807.06732
5. Lu J, Sibai H, Fabry E, Forsyth D (2017) No need to worry about adversarial examples in object detection in autonomous vehicles. arXiv preprint. arXiv:1707.03501
6. Kurakin A, Goodfellow I, Bengio S (2016) Adversarial examples in the physical world. arXiv preprint. arXiv:1607.02533
7. Zhao Z, Dua D, Singh S (2017) Generating natural adversarial examples. arXiv preprint. arXiv:1710.11342
8. Hendrycks D, Zhao K, Basart S, Steinhardt J, Song D (2021) Natural adversarial examples. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 15262–15271
9. Moosavi-Dezfooli SM, Fawzi A, Frossard P (2016) Deepfool: a simple and accurate method to fool deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2574–2582
10. Carlini N, Wagner D (2017) Towards evaluating the robustness of neural networks. In: 2017 IEEE symposium on security and privacy (sp), IEEE, pp 39–57
11. Chen J, Jordan MI, Wainwright MJ (2019) Hopskipjumpattack: a query-efficient decision-based attack. arXiv preprint, vol 3. arXiv:1904.02144
12. Li X, Li J, Dai T, Shi J, Zhu J, Hu X (2021) Rethinking natural adversarial examples for classification models. arXiv preprint. arXiv:2102.11731
13. Engstrom L, Tran B, Tsipras D, Schmidt L, Madry A (2019) Exploring the landscape of spatial robustness. In: International conference on machine learning, pp 1802–1811
14. Athalye A,n Engstrom L, Ilyas A, Kwok K (2017) Synthesizing robust adversarial examples. arXiv preprint. arXiv:1707.07397
15. Tramèr F, Kurakin A, Papernot N, Goodfellow I, Boneh D, McDaniel P (2017) Ensemble adversarial training: attacks and defenses. arXiv preprint. arXiv:1705.07204
16. Athalye A, Carlini N, Wagner D (2018) Obfuscated gradients give a false sense of security: circumventing defenses to

adversarial examples. In: International conference on machine learning, PMLR, pp 274–283
17. Carlini N, Athalye A, Papernot N, Brendel W, Rauber J, Tsipras D, Goodfellow I, Madry A, Kurakin A (2019) On evaluating adversarial robustness. arXiv preprint. arXiv:1902.06705
18. Fezza SA, Bakhti Y, Hamidouche W, Déforges O (2019) Perceptual evaluation of adversarial attacks for cnn-based image classification. In: 2019 Eleventh international conference on quality of multimedia experience (QoMEX), IEEE, pp 1–6
19. Jordan M, Manoj N, Goel S, Dimakis AG (2019) Quantifying perceptual distortion of adversarial examples. arXiv preprint. arXiv:1902.08265
20. Jefferson B, Marrero CO (2019) Robustness metrics for real-world adversarial examples. arXiv preprint. arXiv:1911.10435
21. Sara U, Akter M, Uddin MS (2019) Image quality assessment through fsim, ssim, mse and psnr–a comparative study. J Comput Commun 7(3):8–18
22. Hosseini H, Xiao B, Poovendran R (2017) Google's cloud vision api is not robust to noise. In: 2017 16th IEEE international conference on machine learning and applications (ICMLA), IEEE, pp 101–105
23. Pitropakis N, Panaousis E, Giannetsos T, Anastasiadis E, Loukas G (2019) A taxonomy and survey of attacks against machine learning. Comput Sci Rev 34:100199
24. Directive WF (2003) Common implementation strategy for the water framework directive (2000/60/ec). Guidance document (7)
25. Ruiz-Santaquiteria J, Bueno G, Deniz O, Vallez N, Cristobal G (2020) Semantic versus instance segmentation in microscopic algae detection. Eng Appl Artif Intell 87:103271
26. Redmon J, Farhadi A (2017) Yolo9000: better, faster, stronger. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7263–7271
27. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L (2009) Imagenet: a large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition, IEEE, pp 248–255
28. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A (2015) Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1–9
29. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv preprint. arXiv:1409.1556
30. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
31. Redmon J, Divvala S, Girshick R, Farhadi A (2016) You only look once: unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 779–788
32. Chollet F (2017) Xception: deep learning with depthwise separable convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1251–1258
33. Wang Z, Bovik AC, Lu L (2002) Why is image quality assessment so difficult? In: 2002 IEEE international conference on acoustics, speech, and signal processing, vol 4, IEEE, pp IV–3313
34. Eskicioglu AM, Fisher PS (1995) Image quality measures and their performance. IEEE Trans Commun 43(12):2959–2965
35. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP (2004) Image quality assessment: from error visibility to structural similarity. IEEE Trans Image Process 13(4):600–612
36. Sheikh HR, Bovik AC, De Veciana G (2005) An information fidelity criterion for image quality assessment using natural scene statistics. IEEE Trans Image Process 14(12):2117–2128
37. Sheikh HR, Bovik AC (2006) Image information and visual quality. IEEE Trans Image Process 15(2):430–444
38. Larson EC, Chandler DM (2010) Most apparent distortion: full-reference image quality assessment and the role of strategy. J Electron Imaging 19(1):11006