How fake is my image? Evaluation of Generative Adversarial Networks

Marta Marques
marta.p.marques@tecnico.ulisboa.pt
Instituto Superior Técnico
Lisbon, Portugal

Abstract— **Generative machine learning is a very recent field and has been proven very successful in many multimedia applications. One that has had much impact is the creation of fake images or video clips which can challenge the human perceived notion of reality. Generative Adversarial Networks (GANS) have risen in popularity among the generative models that can learn high-dimensional distributions of data (i.e. the manifold of the entire set of natural images). And have been applied successfully to several tasks such as image compression. Naturally, for these solutions is important to assess the perceptual quality of generative images which will be shown to the user. However, it was recently shown that the typical objective Image Quality Assessment (IQA) metrics have proven to be inefficient at evaluating the perceptual quality of GAN produced images. Also, other quality factors play a role besides fidelity, such as the naturalness/fakeness of the image. Here is presented a subjective and objective study of image quality of GAN created images.**

Keywords—: Image Quality Assessment, Generative Adversarial Networks, Machine Learning, Deep Neural Networks

I. Introduction

Since the introduction of GANs several developments were made in the generation of high-quality images. The initial promising results of the adversarial networks caught the attention of several deep learning researchers, creating an explosion of studies on GANs in the following years. These studies are focused mainly on improving the generated image quality and the stabilization of the training process. Nowadays, many variants of the original GAN have been proposed and can generate high quality attracting images. These GANs have been applied to several image processing problems such as text-to-image or image-to-image translation, image compression, super-resolution, denoising, and other realistic natural image generation applications.

An important part of this type of generative algorithms (which are applied to many computer vision and image processing tasks) is the capability of measuring the perceptual quality of the generated images. The conventional metrics used to quantify image quality are quite ineffective when applied to the content generated by GANs, mainly because generative models can produce images that appear realistic and attractive but do not match when pixel-based comparisons are made. This behavior occurs often in image compression, super-resolution or denoising GAN based systems which have shown superior perceived quality but low objective quality when popular full-reference quality metrics such as PSNR and SSIM are used. On the other hand, no-reference quality metrics do not consider the original information and thus are missing valuable (and important) information.

On the other hand, the images generated by GANs can present some artefacts that show evidence that was synthetically generated (fake) very different from the usual artefacts created by an image compression solution. Typically, these artefacts are not considered in traditional image quality assessment methods. This motivates the need for new image quality subjective assessment experiments (and perhaps methodologies) as well as objective quality metrics suitable to GAN based solutions.

II. Neural Networks: Foundations and Architectures

The idea of having a machine mimicking the way a human brain solves some real-life tasks motivated the creation of biological brain-inspired Machine Learning (ML) models, such as Neural Networks (NN). In a very simplified way, a NN is an interconnected set of artificial neurons. These neurons are mathematical models composed of a set of weights (w_i), a bias (b) and an activation function (non-linear function) that apply the transformation to the input. In a NN the neurons are organized in layers. The neurons in one given layer connected to the neurons in the following layer. The weights of an NN are tweaked in a process called training. In the context of image processing tasks, the training is typically supervised, which means that the ground-truth is provided, and the goal is to minimize the error between the predictions made by the network and the ground truth. In this context, a loss function is used to measure the prediction error, which may vary depending on the problem at hand.

Nowadays, there is a large variety of Neural Networks with different types of layers, topologies, and purposes. Considering that we are focused on GAN-based solutions the main types of NN identified are: Convolutional Neural Networks (CNN), Autoencoder and GAN.

A. Convolutional Neural Networks

CNN are feedforward neural networks that, contrarily to regular neural networks, take advantage of the spatial dependence of its inputs leading to more efficient feature extraction and fewer parameters. The architecture of a CNN may vary greatly depending on the type of problem at hand, but typically there are two key building blocks: convolutional layers and pooling layers.

Convolutional layers are composed of a set of filters or kernels. A kernel is an array of weights, learned through the typical training process, and represents some characteristic of the data relevant. These filters can be 1,2 or 3-dimensional arrays depending on the input's dimension. In this type of layer, the output is computed by sliding the kernels over the input data and performing a dot product. When designing a CNN are a set of

The pooling layers follow the convolutional layers and serve the purpose of downsampling the output of the previous layer.

The need to downsample the output of the convolutional layers comes from the fact that chaining convolutional layers greatly increases the number of parameters in the NN. These pooling layers work with sliding filters, like the convolutional layers. The most common pooling layer is the max-pooling layer, that as the name suggests, returns the maximum value of the inputs.

B. Autoencoders

Autoencoders (AE) are a type of unsupervised learning algorithm. More specifically are designed to have the output approximate the input with some constrains or restrictions, most often requiring the input to be represented also with lower dimensionality. These restrictions built into the neural network allow to represent the input data in some compressed way (as the name auto-encoder suggests) and in this process learn the most important features of the input data. This compressed representation (also called code) lies on some latent space. During the training of the auto-encoder, the encoder and decoder work together trying to recreate as closely as possible the input but with certain restrictions (typically in the architecture design) to prevent a simple copy of the data along the network. The AE performance is heavily dependent on the type of data that is trained on. That is to say that the AE is a data specific solution.

C. Generative Adversarial Network

Generative Adversarial Autoencoders are a type of generative model proposed in 2014 by Ian Goodfellow *et al*. [1]. This model, unlike other generative models (e.g. the variational autoencoder), does not estimates explicitly a probability distribution but learns it implicitly during training from examples.

GANs are composed of two separable neural networks, the Generator (G(x)) and the Discriminator (D(x)) and exploit a game-theory approach where the two NNs compete against each other as adversaries. Moreover, the goal of the Generator is to create synthetic (fake) data that resembles as closely as possible to real data and thus, fool the Discriminator, whose goal is to distinguish between fake data and real data.

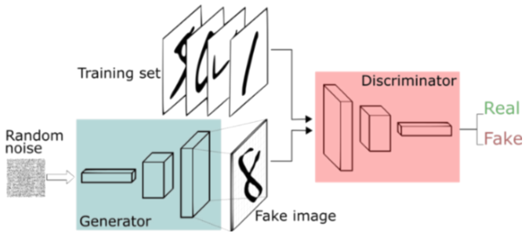


Fig. 1. Architecture of a Generative Adversarial Network.

In practice, the Generator receives samples from a simple distribution (e.g. random noise) which is processed by a sequence of neural network layers, and the output corresponds to data that resembles the data from the training set at least in a semantic way. However, the Generator is blind to the training dataset, being dependent on the Discriminator to guide its learning process. The Discriminator receives data samples and outputs probabilities that represents if the sample belongs to the training dataset or not. This means that the discriminator evaluates the authenticity of images or how close the images produced by the generator are close the images of the training dataset. When the generator reaches a high level of performance (this means produces realistic looking images), the Discriminator is unable to distinguish between real and fake data, which means $D(x) = 0.5$ for every sample. Thus, Generator goal is to produce images that are considered realistic (i.e. act as a forger without being caught) and the discriminator goal is to assess if these images are fake or not (i.e. act as the police to detect forgeries), as depicted in Fig. 1.

The random noise (z) is the input of the generator that drives the creation of fake image. This fake image and the training set (real) images (x) are fed to the Discriminator, that classifies each image as being either fake or real. This design of a GAN imposes that the cost function Ex[log(D(x))] + E_z[log(1-D(G(z))] (**1**) to be a minimax game, where the Discriminator try to maximize the cost function and the generator to minimize it. Where E_x and E_z are the expected values over all real date and over all random inputs to the Generator, respectively.

$$E_x[\log(D(x))] + E_z[\log(1-D(G(z))] \qquad (1)$$

III. Relevant Image Coding Deep Learning Based Solutions

GAN can be applied to a variety of image processing problems, such as image compression, super-resolution, and artifact removal. Three of those solutions were selected to be reviewed. The solutions were selected due to its performance and overall popularity and will serve as a starting point for the remainder of this study.

A. Generative Adversarial Networks for Extreme Learned Image Compression

This solution [2] targets extremely low bitrates, where is hard to preserve many visual elements of an image with high fidelity. The solution described in this section proposes to generate artificial elements and can be classified as extreme compression, where the pixel-wise preservation becomes less important when compared to the global structure and the semantic meaning of the image.

The proposed solutions optimize the image coding process beyond the usual conventional image quality metrics used in the training process, with a loss function that includes an adversarial loss term. This framework has two operation modes, namely:

- Generative Compression (GC): This mode of operation exploits the generative capabilities of GANs to aid in image compression, preserving the content as much as possible. This solution is similar to the vanilla GAN described but introduces a new loss function and instead of sampling from a random noise distribution, the generator (decoder) uses the encoded image as its input.
- Selective Generative Compression (SC): This mode of operation preserves some elements of the original image with high fidelity and other elements are only generated based on semantic information about these elements. The SC uses some additional information about the image, namely, its semantic label map, which can be seen as a Conditional GAN (cGAN) [3]. This solution is particularly suitable in scenarios where a part of the image is perceptually very relevant, while other parts are less relevant.

B. Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network

In 2017, Ledig *et al.* [4] proposed a generative adversarial network for image super-resolution (SRGAN) that can up-sample images up to 4x using a single low-resolution image. The introduction of the GAN adversarial loss makes the higher resolution (HR) images richer in high-frequency details, when compared to other super resolution (SR) algorithms that use MSE in the loss function. When the MSE loss function is used, pixel-wise similarities between the ground-truth and the SR image are accounted and thus, high PSNR quality scores are obtained; however, this often results in perceptually unsatisfying HR images.

Also, this work introduces a perceptual loss function that balances the generation of content with the preservation of original LR image content. This new training objective coupled with a ResNet inspired architecture creates a solution that outperforms previous state-of-the-art models.

In this work, two different models for the super-resolution problem are proposed:

- SRResNet model: an application of a ResNet [5] trained with the MSE as its loss function. This network is composed of 16 residual blocks. The fact that the model is optimized for MSE makes it more sensible to pixel changes and invariant to perceptual changes. This model sets a new state of the art in image SR in PSNR and SSIM measures.
- SRGAN model: GAN based solution composed of the usual Generator-Discriminator pair. The generator creates reconstructed images very similar to the real images and thus difficult to classify by the discriminator, which attempts to distinguish between the real images and the reconstructed high-resolution images created by the generator.

C. Deep Universal Generative Adversarial Compression Artifact Removal

Galteri *et al.* [6] proposed, in 2019, a compression artifact removal solution that is efficient even in highly degraded images and works in scenarios where the encoding parameters are unknown. This is different from previous deep learning-based methods which assume that the encoding parameters are known, namely the compression factor QF. To achieve this objective, the authors introduce an ensemble of deep convolutional residual networks, trained for different compression quality factors (QF), coupled with a quality predictor model. Each model its trained for different compression qualities, i.e. from lower to higher compression rates, allowing for a tailored image reconstruction.

The core of this solution is a neural network that can be trained with direct supervision or with adversarial training. The authors explore several configurations for this model, namely using the GAN framework and balancing the adversarial loss with a content loss or using direct supervision with MSE and SSIM. Regarding the content loss several alternatives were evaluated, such as the MSE, SSIM, and VGG19 [7].

The proposed solution introduces an ensemble of GAN models (generators), which are convolutional networks trained with decoded images compressed with different quantization factors QF. The input compressed image (I^C) is passed to a compression quality predictor NN that estimates the image QF. With this estimation the QF switcher passes the I^C to the better suited GAN and the reconstructed I^R image is obtained.

IV. GAN-BASED IMAGE PROCESSING: SUBJECTIVE QUALITY EVALUATION

Having explored the literature and established the main objectives of the subjective quality assessment the following steps are to determine the conditions under which the test will be executed. That is the methodologies applied in the assessment.

A. Test Material and Preparation

To design a subjective assessment, it is important to define what will constitute the test material. That is to say, which images will compose the study, how will the images be shown and what preparations need to be done beforehand.

1) Dataset selection

The JPEG AI dataset [8] was used for this subjective quality evaluation seeing that it was designed to evaluate the performance of state-of-the-art learning-based image coding solutions. Due to its nature, this dataset is divided into training, validation, and test subsets. The test material used in this assessment belonged to the test subset since it provides a diverse and well-balanced set of images that allows the evaluation to be representative.

To keep the test session restrained to less than 60 minutes from the 16 total images that compose this test set only 8 are used in the subjective assessment. The selection of the 8 test images is done by visual inspection and considering the need for diversity of contents. The images are also selected to range from less to more complex when it comes to high-frequency components, color saturation, etc. Leading to the test set described in TABLE I.

TABLE I. IMAGES FROM JPEG AI TEST DATASET SELECTED FOR SUBJECTIVE ASSESSMENT

Image ID	Code Name	Image ID	Code Name
00002	Racing car	00008	Transmission towers
00004	Rotunda of Mosta	00009	Port
00005	Las Vegas sign	00010	Curiosity Rover
00006	Train	00012	Woman

2) Display Conditions

Normally the subjective assessment would be done in a controlled environment where all subjects would visualize the images in the same monitor with the same configurations. However, due to the limitations imposed by the global pandemic, the approach shifted in the direction of an online crowdsourcing survey.

With the assessment done in different devices, there are consequently different viewing conditions of the test material. This creates the need to control as much as possible the conditions under which the survey is taken. In order to do so, the app developed to employ subjective assessment imposes restrictions regarding the resolution and size of the display. Particularly, it enforces a resolution of at least 1920x1080 and a display size of at least 13 inches.

3) Content Preparation

Considering that this assessment follows a pairwise comparison design, meaning that two images need to be displayed side-by-side, it is important to determine the target size for the test material. With the previously mentioned target resolution of 1920x1080 the selected desired resolution for the test images is 940x880. This is done by making crops over the test images while accounting for the need to preserve the relevant elements of each image.

B. GAN-based Image Processing Solutions and Benchmark

The next set in the construction of subjective assessment is the creation of the test material. To create the test images are a set of procedure that were followed. Namely regarding preprocessing of the images, the training of the model, the postprocessing of the images and other related aspects.

Regarding the GAN-based solutions used in this section, they were selected having in mind the importance of the task performed, the performance, and the availability of software. For each solution the objective is to obtain 3 levels of image quality to evaluate the impact the perceptual impacts of the artefacts generated by GAN based solutions.

1) Enhanced Super-Resolution Generative Adversarial Networks (ESRGAN)

We could not find an official publicly available software for the previously describe GAN-based solution that performed SR. Considering other choices, it was selected the ESRGAN proposed by Xintao Wang *et al.* in [9] for which an implementation was available. The ESRGAN model is an improved version of the solution described in Section III.A and introduces some new tools. Among those, a Residual-in-Residual Dense Block (RDDB) with an improved discriminator network based on Relativistic average GAN (RaGAN) [10]. Regarding the generator used, it follows an SRResNet [4] architecture and had 32 layers of RRDB blocks.

Three variants of the proposed solution were trained with the goal of obtaining three distinct level of artifacts caused by the content generation capabilities of the network. This is done by changing the weight (λ) of the adversarial loss (L_G^{Ra}) in the loss function of the model $\boldsymbol{LG = L_{percep} + \lambda L_G^{Ra} + \eta L_1}$ **(2)**.

$$\boldsymbol{L_G = L_{percep} + \lambda L_G^{Ra} + \eta L_1} \quad (2)$$

This means that the adversarial loss can have more or less importance, and thus may lead to images with higher apparent quality but less real, i.e. more fake. The different ESRGAN models can be defined as ESRGAN Lo for $\lambda = 5 \times 10^{-2}$, ESRGAN M for $\lambda = 10^{-2}$ and ESRGAN Hi for $\lambda = 5 \times 10^{-3}$.

The ESRGAN Lo and Mi models were trained following methodology presented by Xintao Wang *et al.* only deviating from it in the mini-batch size that was set to 4 instead of the recommended 16 due to hardware limitations. The ESRGAN Hi model was a provided pretrained model that follow the suggested training procedure.

After the models were trained, the next step is to obtain the super resolution images for the test set. To perform the super resolution task on the test set, low-resolution (LR) images are first obtained. These LR images are obtained as recommended with a Matlab function that performs resize using a bicubic kernel. More specifically, an upscaling factor of 4 is used. It is important to note that the crops mentioned in Section A.3) are done after the super resolution image is obtained and not before.

2) High-Fidelity Generative Image Compression (HiFiC)

The compression solution presented in Section III.A., also doesn't have a public available software. Moreover, recently a GAN-based image compression solution was proposed was proposed by Mentzer *et al.* [11] which is considered more efficient and is nowadays recognized as highly efficient. Having the official implementation of this improved GAN-based compression solution we use it to obtain images for the subjective assessment study.

This new image compression solution uses some of the same tools present in SC mode of the older model, such as, the use of a Conditional GAN. However, the HiFiC model introduces a new objective function (3) that targets the minimization of the rate-distortion trade-off and an architecture that improves the training.

$$\mathcal{L}_{EGP} = \mathbb{E}_{x\sim px}\left[\lambda r(y) + d(x,x') - \beta\log\left(D(x',y)\right)\right] \quad (3)$$

Along with the image compression solution, 3 pre-trained models are made available for 3 distinct target bitrates, more specifically they are described as HiFiC Lo for $r_t = 0.14$ bpp, HiFiC Mi for $r_t = 0.3$ bpp and HiFiC Hi for $r_t = 0.14$ bpp.

For the subjective assessment experiments, it was crucial that these 3 models presented visible differences in quality, from minor to severe degradations. However, after close inspection of the images produced by these pretrained models it was clear that the difference in perceptual quality between the HiFiC Mi and HiFiC Hi models was almost nonexistent. Moreover, the pretrained models lead to images with very few artifacts. Therefore, a new HiFiC model was trained, which is more restrictive in terms of bitrate, in this case, the target rate was set to 0.06 bpp.

The training of this new model follows the methodology proposed by the authors of the proposed image compression solution. The training has two steps: i) training with a simple objective function where distortion is given by (4) and ii) training with discriminator.

$$d = k_M \mathrm{MSE} + \mathrm{k_P LPIPS} \quad (4)$$

Where k_M and k_P are hyperparameters that balance the weight of each component. These hyperparameters are fixed in the experiments described by the authors and therefore, will also be fixed in the same way in our training.

The training dataset used was the same as the experiments described in the paper, that is the Coco 2014 [12] dataset. It is also important to note that besides the configurations changed in order to set the desired target rate it was also changed some configurations in the *model.py* file due to hardware limitations.

After training the model we obtain the images of interest for the test, however it was clear that the network was unable to reach the desired birate. This could be fixed by studying the impact of the hyperparameters in the training of the model, but the images resulting of this model were already different from those obtained with other models, and as such were used in the study. In summary, the HiFiC models used can be defined as HiFiC Lo for $r_t = 0.06$ bpp, HiFiC Mi for $r_t = 0.14$ bpp (pretrained) and HiFiC Hi for $r_t = 0.3$ bpp (pretrained).

3) Artifact removal GAN (ArNet)

Once again, the model described for the artefact removal task in Section III.C did not have publicly available software.

After considering several models with public software available, the selected solution for artefact removal was ArNet [13], a more recent and efficient model which had a public available implementation along with a pretrained model. The main novelty of ArNet is the NoGAN [14] technique, a new methodology to train the GANs. The NoGAN training pretrains the generator and discriminator with straightforward, fast, and reliable conventional loss functions, which are then trained together in a normal GAN setting. However, the number of iterations that the generator and discriminator are trained is limited, resulting in a model that is faster to train and with higher quality (less artefacts).

To perform subjective assessment, 3 distinct levels of quality should be obtained as in the other GAN based solutions. In this case, the network used to obtain the images was the pretrained available network, and to achieve different levels of quality, the target JPEG decoded quality was varied and consequently the difficulty of the artifact removal task. In this case, the quality factor was changed. The quality (QF) values used for the experiment and corresponding nomenclature used in the experiment are the following: ArNet Lo for $QF = 7$, ArNet Mi for $QF = 14$ and ArNet Hi for $QF = 21$.

The JPEG compressed images were obtained using the Python PIL module (save function). Then those images are passed through the NN. Before being added to the image test set the ArNet generated contents are cropped to fit the testing environment.

C. Subjective Test Methodology

Some methodologies employed in the test still need to be defined, e.g., the screening of the test material, the grading method used and the environment developed to deploy this test. The choices made have into consideration the ITU-R Recommendation BT.500-14 [15] and the insights provided by other subjective assessments done with crowdsourcing [16].

1) Design of the subjective test

The forced choice method used in this subjective test was selected since indirect scaling methods have a high discriminatory power. Moreover, the GAN-models introduce small changes in the perceived quality which amplifies the need to employ a method that can translate these differences into measurements. The force choice method is also easier for the subjects to perform comparisons which was another concern due to the fact that the test was made via web application and consequently without in person interaction.

As previously mentioned, the test material consisted of 8 different pieces of content or reference images (n), and 10 conditions representing different quality levels (k) . Moreover, these 10 conditions correspond to 3 levels of quality for each task (compression, super resolution, and artefact removal) and the original image. To have a complete design of the subjective test, that is every image is compare with every other image, each subject would need to perform $C_2^{n\times k} = 3160$ evaluations. Which would create test with a duration that is not reasonable and to avoid this problem the design chosen is incomplete. Moreover, for each refence image all pairs are considered, amounting to $n \times C_2^k = 360$ comparisons. Then it is also added some cross-content pairs to enable the scaling of the image quality scores between different pieces of content. These cross-content pairs where only the reference images with adjacent image indexes for the same task and quality level, which results in additional 72 pairs. Lastly, it was included a benchmark comparison for image compression which translates to other 24 comparisons, leading to a total of 456 comparisons.

The main disadvantage of using FC method is the complexity of the data processing, since the judgements do not translate directly the score of a given image and the incomplete design of the test creates the need to extrapolate from the given judgements.

Regarding the duration of the subjective test assessment, the ITU-R Recommendation BT.500-14 [15] we dived the test into two sessions of maximum 30 minutes each. This is done to avoid that the subjects are too fatigued to make accurate judgements. When it comes to the environment in which the test is performed due to the fact it was done via web application it was not possible to maintain a controlled viewing experience for each subject, however some limitations were made to the display in which the test is done. Moreover, the screen resolution and size were restricted to a minimum of 1080x1920 and 13 inches, respectively.

2) Implementation of the subjective test environment

The test environment was a web application that can be seen as the combination of 3 main functional components: HTTP server, database, and graphical user interface (GUI). These modules work together as depicted in Fig. 2. Simply put the GUI provides the environment and information that the user needs to perform the test, the data provided by the user in the GUI is passed to the server that connects to the database to store the data and serve new HTTP pages accordingly.

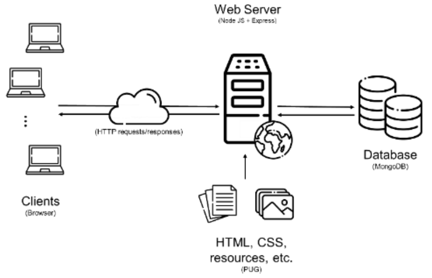
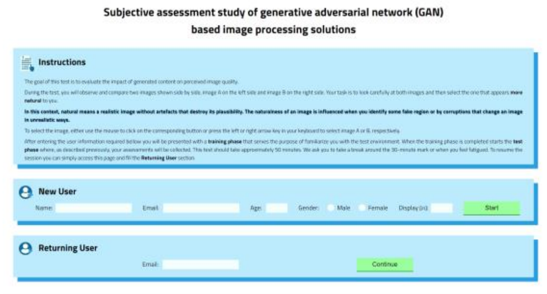


Fig. 2. Schematic of the architecture of the subjective test web application

In the implementation of this web application we opted to develop the server using Node JS [17] and the Express framework [18]. In conjunction with the Node JS a simple MongoDB database [19] was deployed. This DB contained 3 collections, the training collection that served the purpose of organizing the images in the test set and the pairs that would be shown during the evaluation. The session collection that stored the information regarding each subject. Which includes a set of personal details and viewing conditions, as well as the specific order in which the user is shown the test pairs and its corresponding decision accompanied of the date and time in which it was made. And lastly, the training collection, that organized the images and information relevant to the training phase.

Regarding the user interface it was developed using a template engine called PUG [20]. This allowed the development of dynamic HTML pages with complete abstraction from the content that would be display. To help organize and style the resources in the web page layouts it was used Cascading Style Sheets (CSS).

The typical flow of the web application is as follows: the user connects to the server requesting the home page of the website, depicted in Fig. 3.



Subjective assessment study of generative adversarial network (GAN) based image processing solutions

Instructions

The goal of this test is to evaluate the impact of generated content on perceived image quality.

During the test, you will observe and compare two images shown side by side, image A on the left side and image B on the right side. Your task is to look carefully at both images and then select the one that appears **more natural** to you.

In this context, natural means a realistic image without artefacts that destroy its plausibility. The naturalness of an image is influenced when you identify some fake region or by corruptions that change an image in unrealistic ways.

To select the image, either use the mouse to click on the corresponding button or press the left or right arrow key in your keyboard to select image A or B, respectively.

After entering the user information required below you will be presented with a **training phase** that serves the purpose of familiarize you with the test environment. When the training phase is completed starts the **test phase** where, as described previously, your assessments will be collected. This test should take approximately 50 minutes. We ask you to take a break around the 30-minute mark or when you feel fatigued. To resume the session you can simply access this page and fill the **Returning User** section.

New User

Name: Email: Age: Gender: ○ Male ○ Female Display (in): Start

Returning User

Email: Continue

Fig. 3. Homepage of the subjective test web application

The server responds with a HTML page containing a brief context and instructions for the assessment as well as two option to start a testing session. The first applies to new users that have not yet started the test, in this case the user is asked to provide some information. The user then must pass the resolution (at least 1080x1920p) and display (at least 13 inches) tests enforced by the application in order to start the test. Once these tests are valid the user is added to the database and starts what is called the training phase where a sequence of images are shown with the purpose of familiarize the observer with the type of test material that will be shown and the environment in which the test will be performed in.

If the user has already started the test assessment and is returning to the web application for the second session, then in homepage is asked to identify itself with the email and the server responds with the resuming of the subjective test.

In the test page, shown in Fig. 4, the user is shown the 2 images to be compared side by side and can input its decision by either clicking on the corresponding button or by pressing the right or left arrows on the keyboard. Corresponding left arrow key to the option “A more natural than B” and the right arrow key to the “B more natural than A” option. And the judgment information is stored in the corresponding session document.

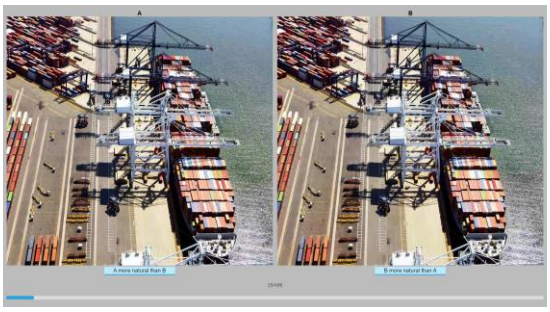


Fig. 4.Test environment of the subjective test. On the left selection screen and on the right the end of test screen.

D. Data Processing: converting pairwise-comparison to psychophysical scores

The assessment produces raw data that consist of a binary array of decisions per participant. This data by itself does not provide any intuitive notions of the quality of the images. So, there is the need to process the raw data in order to extract from it relevant information.

But firstly, is important to identify unreliable observers, thus invalidating its judgements from further examination. And then perform the transformation of the binary information made at the pair level to the quality scores that represent each image in the study.

1) Outlier Detection

Outlier detection is a procedure that can have different natures. One is to compare the responses given by a certain observer with the rest of the observers in the study. This method serves the purpose of establishing if there is agreement between the subjects in the study, i.e., the same criterion of assessment is used. The other is to evaluate the individual fidelity of the subjects.

Due to the crowdsourcing nature of this assessment is crucial to evaluate the fidelity of the judgements made by the participants. To verify the reliability of a given user we use the transitivity rule [21], [22]. To quantify the impact of this occurrences is employed a metric called Transitivity Satisfaction Rate (R_i) defined by the frequency in which this rule was satisfied (5**).**

To compute R_i firstly was constructed an adjacency matrix from the raw data which can be interpreted as directed graphs which will facilitate the computation of R_i.when applying the *NetworkX* python package [23].

$$R_i = 1 - \frac{M_i}{N_i} \tag{5}$$

Any subject that presented a transitivity satisfaction rate lower than 0.8 would be considered unreliable and consequently an outlier being then dropped from the study. No users were dropped since all complied with the requirements established.

2) Quality Scores

Having validated the consistency of judgements made by the parties in the study, there is the need to convert the raw binary data into quality scores. In this experiment first is obtained a winning frequency matrix and later is applied the Bradley-Terry Model [24]. The wining frequency is a simple method that provides good insight to the results in the study, however it presents it in an ordinal scale lacking in the ability to capture the real magnitude of difference in quality between the images. This problem is amplified by the incomplete design of the test. To better analyze the data is then used the Bradley-Terry Model, that calculates the probability of winning of any given element in a pair comparison. There are several variations of the Bradley-Terry model, in our analysis we use the implementation provided by the python library *Choix* [25].

3) Test Resukts Analysis

In this study participated a total of 31 assessors. Some observers were experts while others were non-experts, however none had any prior contact with the experiment’s test material. From those 31 subjects, 21 were male and 10 were female. Regarding the ages of the observers, they were comprised between 20 and 55. The viewing conditions were varied in this study, however most observers performed the test in displays with at least 20 inches.

Every observer was asked to perform 456 comparisons resulting in a total of 14 136 judgements. Due to its length the test was divided in two sessions. These 2 sessions had durations comprised between 20 to 30 minutes depending on the user. Resulting in an average total test duration of approximately 58 minutes. From the response times records gathered most subjects take on average less than 7 seconds to make a judgment.

Regarding the judgments raw data, the first analysis is to evaluate if there was preference for some of the reference images. This is done by computing the winning frequency by content. That is to say that the vote count is performed by grouping the assessments by reference image. If there is no bias towards the contents of the images, then the winning frequency would be of 0.5 for every reference image. In Fig. **5** are shown the results obtained for this experiment, and, overall, the images fall very close to the 0.5 winning frequency that implies no preference for specific contents.

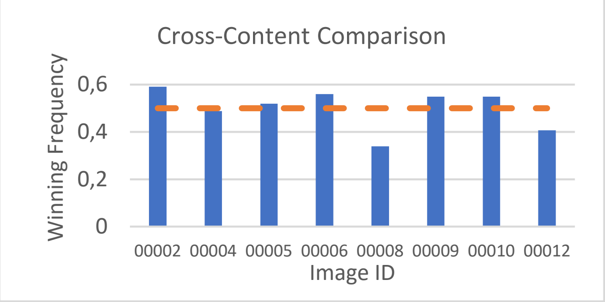


Fig. 5. Comparison of the winning frequencies by image content.

It was also computed the winning frequencies for the test pairs relating to the comparison between HiFiC and HEVC for similar compression factors. In *Fig.* **6** are depicted these results from which is visible that the HiFiC model was largely preferred when compared with HEVC – Intra for the same bitrate. Leading to the conclusion that the HiFiC solution produces images that are considered *more natural* than one of the best conventional image codecs.

With the decrease in bitrate (*Fig.* 6 right to left) there is an increase in preference for the images produced by HiFiC. This highlights the fact that the GAN-based model excels in more challenging tasks, where it utilizes more of its generative capabilities. Also showing that, in these harder compression scenarios, the artefacts introduced by the GAN do not worsen greatly the *naturalness* of the image.

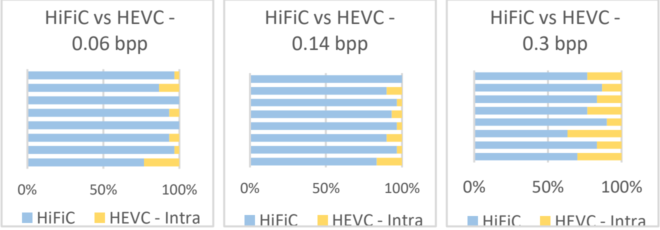


Fig. 6. Comparison of the winning frequency between HiFiC and HEVC-Intra with similar bitrates. Left: 0.06 bpp. Centre: 0.14 bpp. Right: 0.3 bpp.

A similar analysis was done for each solution comparing the different models with each other and the original reference image, which is presented in Fig. **7**. The behavior displayed by all the solutions reflect the expected increasing preference by the GAN-created images with the increasing ease of the task as well as the obvious inclination for the original reference image.

For the ESRGAN solution Fig. **7** (left) there is a very constant increase in winning frequency, approximately doubling every model. Since the main difference between these models (Lo, Mi, Hi) was the tweaking of the hyperparameter λ that controls the weight of the generative content in the training process, is clear that the *naturalness* of the produced images was affected by the artefacts introduced by the GAN.

Regarding the HiFiC solution the Fig. **7** (center) shows that although it follows the expected increase in winning frequency with the increase of the target bitrate for each model, the model HiFiC Mi and HiFiC Hi are more comparable with each other than with the HiFiC Lo model or the reference original image. Which in combination with an average winning frequency of less than 40% for the reference content when compared with HiFiC images, showcases the great performance of this solution

Lastly, the ArNet solution winning frequency distribution presented in Fig. **7** (right) highlights the poor performance of the Lo model. Which was anticipated since the model was trained for less severe degradations then the one used in ArNet Lo.

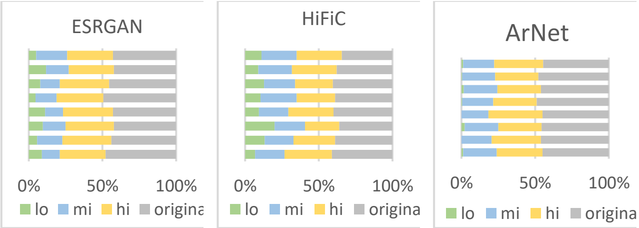


Fig. 7. Comparison of the winning frequency of each model by solution. Left: ESRGAN. Centre:HiFiC. Right: ArNet.

Besides the winning frequency analysis, it was also performed a scaling of the raw data using the *Choix* python library. More specifically, we obtained a penalised Maximum Likelihood Estimation (MLE) using the Bradley-Terry model. These results are presented in Fig. **8** and were grouped by reference images, being also included the average MLE.

Once again it is visible the increase of perceived quality from models Lo to Mi to Hi in all solutions. However, for the specific case of the *Woman* (00012) images obtained with HEVC-Intra coding it is visible that the MLE of the Mi model breaks this tendency since it has a very low MEL value. This comes from the fact that the HEVC codec was only introduced in the assessment by comparisons with HiFiC and had no votes the MLE value.

Analysing Fig. **8**, we can infer that the compression solution HiFiC produced the more *natural* looking images. Being also the more recent and complex solution, this was expected. Once again, the variation in quality between the Mi and Hi models is very small corroborating the analysis made in with winning frequency data.

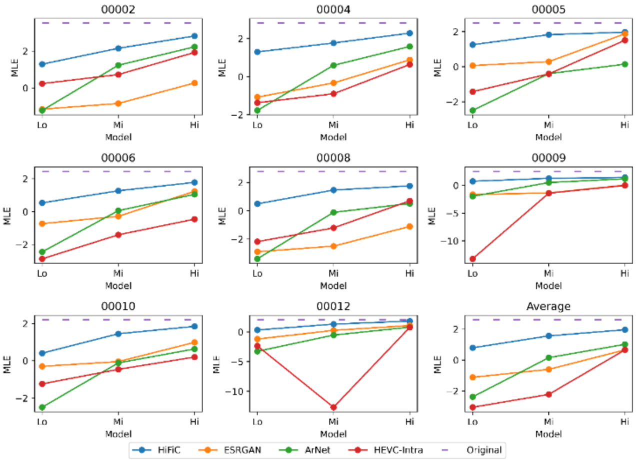


Fig. 8. Comparison between different solutions MLE by image of reference.

The ArNet solution follows the HiFiC model when it comes to the naturalness of its images for the model Mi and Hi. However, for the Lo model the ESRGAN takes second place. The poor quality of images produced by the ArNet Lo was already hinted in the vote count analysis and it is supported in the MLE evaluation.

Although, the ESRGAN has the worst image quality when compared with the other GAN-based solution in this study, is important to note that the 4x super resolution task is rather challenging when compared with the other tasks in this study. In the *Racing car* (00002) and *Transmission Towers* (00008) images the ESRGAN produces very poor images indicating the inability of the model to reconstruct details.

V. Quality Metric performance evaluation

The human eye is the best judge of image quality, however, is not feasible to rely on subjective assessments every time there is a need to evaluate the image quality. This motivates the need to have objective quality metrics that can accurately quantify the image quality. We will use the subjective scores obtained in Section IV.D to evaluate the performance of some objective metrics.

A. *Qaulity Metrics*

The metrics studied in this chapter are divided into two sub-categories: Full-reference and no-reference. Full-reference metrics are metrics that rely on the unaltered image and use it evaluate the degradation on the tested image. On the other hand, no-reference metrics are blind to the expected image.

1) *Full-Reference Metrics*

There are a variety of full-reference metrics well known metrics, in this study we will focus only on the following: Peak Signal-to-Noise-Ratio (PSNR), Structural Similarity Index (SSIM), Multi-Scale Structural Similarity Index (MS-SSIM) and Visual Information Fidelity (VIF).

The PSNR is defined by the ratio of maximum possible pixel value of the image (MAX_I) and the Mean Square Error (MSE), as depicted in (6).

$$\mathrm{PSNR} = 10\log_{10}\left(\frac{MAX_I^2}{MSE}\right) = 20\log_{10}\left(\frac{MAX_I}{\sqrt{MSE}}\right) \qquad (6)$$

Being the MSE computed as presented by (7). Where m and n are the image's width and height respectively. Also, I denotes reference image and I' refers to the altered image.

$$MSE = \frac{1}{m\,n}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}[I(i,j) - I'(i,j)]^2 \qquad (7)$$

Unlike the previously mentioned metric, the SSIM [15] is designed with the human visual system (HVS) in mind. Moreover, this measure explores the structural information of an image. That is, the aspects of an image that define its elements. And can be calculated using (8)

$$SSIM(x,y) = \frac{\left(2\mu_x\mu_y + C_1\right)\left(2\sigma_{xy} + C_2\right)}{\left(\mu_x^2 + \mu_y^2 + C_1\right)\left(\sigma_x^2 + \sigma_y^2 + C_2\right)} \qquad (8)$$

Where x and y denote samples from the tested and reference images, respectively. And μ represents the mean of the signals and σ the variance. Regarding the variables C_1 and C_2, they are constants and can be calculated as shown in (9). Where L is the dynamic range of the pixel values and the default values of constants K_1 and K_2 are 0.01 and 0.03 respectively.

$$C_1 = (K_1L)^2$$
$$C_2 = (K_2L)^2 \text{ and } C_3 = C_2/2 \qquad (9)$$

The MS-SSIM [27] provides an improvement over the SSIM metric by computing it on variations of the image resolution and viewing conditions.

The VIF [28] is another metric that explores the characteristics of the HVS in order to better quantify image quality. This metric exploits natural scene statistics (NSS) together with models for the distortion channel and the HVS. It computes the mutual information between the reference image and the reference image as altered by the HVS model to quantify the perceived information that the brain could under ideal conditions obtain. Similarly, calculates the mutual information between the reference image and the image affected by both the HVS model and the channel distortion model.

2) *No-Reference Metrics*

The no-reference images tested in this analysis were Blind/Referenceless Image Spatial QUality Evaluator (BRISQUE), Unified No-reference Image Quality and Uncertainty Evaluator (UNIQUE) and hyperIQA.

The BRISQUE metric employs a natural scene statics model to predict the severity of distortions present in a given image. That is to say that BRISQUE [29] measures the deviation between some local luminance signal's statistics and the expected result using the natural image model.

The UNIQUE [30] metric consists of a DNN trained to infer the image quality. This metric's network has an ResNet architecture and is trained on both synthetically and realistically distorted images.

The hyperIQA [31] is a metric that also consists of a NN. However, the architecture of this NN differs from the simpler UNIQUE network. This NN first gets the semantic features and then makes quality predictions. It also utilizes multi-scale content to better describe the local and global distortions.

B. *Quality Metrics Performance Evaluation Procedure*

In order to evaluate the test set using the metrics mentioned previously we use a variety the python libraries. Moreover, we use the *skimage.metrics* [32] to compute the PSNR and SSIM values. The *Sewar* [33] package to obtain the values for MS-SSIM and VIF metrics.

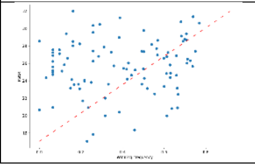
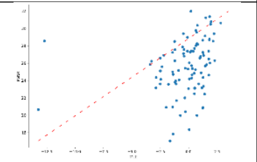
Regarding, the no-reference images the BRISQUE implementation used was the *image-quality* [34] package. The UNIQUE values were computed using the first party software. And for the hyperIQA metric it was also used the official software.

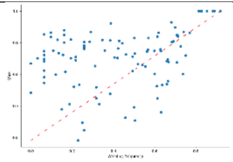
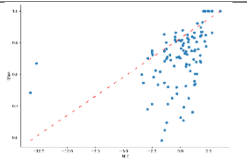
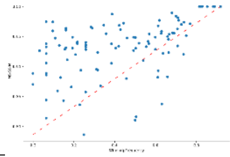
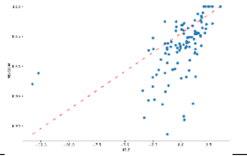
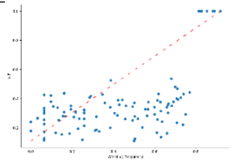
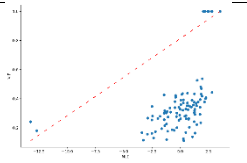
C. *Results and Analysis*

Having obtained all the previously mentioned metric values for all images in the assessment it is performed an analysis that will give insight to the correlation between objective and subjective quality scores.

Starting with the full-reference metrics it was plotted, in TABLE IIErro! A origem da referência não foi encontrada.**,** each metric value as a function of the corresponding both the winning frequency and MLE. These plots can help highlight possible correlations between the objective and subjective scores. Moreover, if the sparsity of the scatter points is greater the correlation between the variable is probably lower.

TABLE II. Wining frequency and MLE plots as function of full-reference metrics.

Metric	Winning Frequency	MLE
PSNR		

Metric	Winning Frequency	MLE
SSIM		
MS-SSIM		
VIF		

From the visual data presented in TABLE II is evident that the PSNR metric does not seem to be correlated to either the winning frequency or the MLE. The SSIM, MS-SSIM and VIF, look to be comparable in terms of sparsity. These metrics are an improvement when compared with the PSNR, however the correlation seems to be only moderate.

The intuitive notions provided by the visual data when studying the correlation between two variables is clearly insufficient and, thus, the analysis continues by computing two different correlation metrics: the Pearson Correlation Coefficient (PCC) and the Spearman's Rank Correlation Coefficient (SRCC).

The PCC is widely used when it comes to quantifying correlations between 2 variables. Its values range from -1 to 1, where -1 signifies to total negative correlation and 1 total positive correlation. Even though SRCC is less used, it has some advantages when compared with the Pearson, namely the is much more resistant to the existence of outliers and data entry.

The values of PCC and SRCC were computed for every full-reference metric and are shown in TABLE III and TABLE IV, respectively. Has suspected, the PSNR has a weak positive correlation to both winning frequency and MLE. Which means it cannot accurately reflect the quality of the GAN-based solutions, which was expected as per the motivation of this assessment.

The SSIM, MS-SSIM and VIF are an improvement over the PSNR being barely moderate correlated to the subjective quality scores.

TABLE III. PCC values for full-reference metrics.

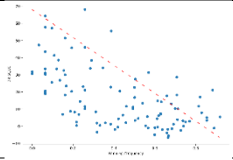
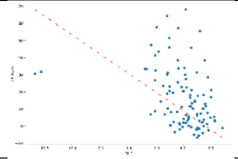
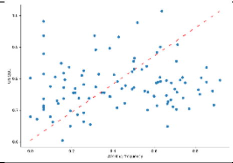
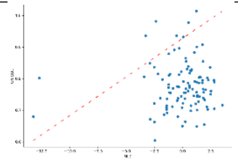
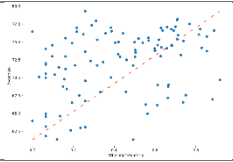
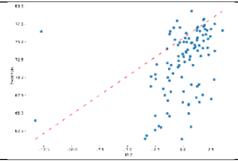
Metric	Winning-Frequency	MLE
PSNR	0.12468232	0.25372442
SSIM	0.33719724	0.40622694
MS-SSIM	0.46789373	0.52074356
VIF	0.54274867	0.46036398

TABLE IV. SRCC values for full-reference metrics.

Metric	Winning-Frequency	MLE
PSNR	0.13491130	0.36734943
SSIM	0.33821214	0.54140478
MS-SSIM	0.51318719	0.67314240
VIF	0.43608015	0.61356022

Regarding the no-reference metrics the methodology was the same. Firstly, are obtained the plots of winning frequency and MLE with each metric. From the graphics shown in TABLE V is evident that UNIQUE metric seems to be uncorrelated to the subjective scores. And BRISQUE and hyperIQA having apparently some correlation with the winning frequency and MLE, having BRISQUE a negative correlation and hyperIQA positive correlation.

TABLE V. Wining frequency and MLE plots as function of full-reference metrics.

Metric	Winning Frequency	MLE
BRISQUE		
UNIQUE		
hyperIQA		

Once again, the procedure is the same has the one used for the full-reference images and are calculated both PCC and SRCC scores. Analyzing the correlation coefficients displayed in TABLE VI and TABLE VIII s clear that BRISQUE is the best no-reference metric to evaluate this type of solutions since is the one with the highest correlation with the subjective scores. Nevertheless, this correlation is only moderate thus not constituting a very reliable assessment.

TABLE VI. PCC values for no-reference metrics.

Metric	Winning-Frequency	MLE
BRISQUE	-0.5857174895	-0.35932435
UNIQUE	0.114379791	-0.103063063
hyperIQA	0.338668255	0.37315125

TABLE VII. SRCC values for no-reference metrics.

Metric	Winning-Frequency	MLE
BRISQUE	-0.58564549	-0.42889149
UNIQUE	0.13559929	-0.12975568
hyperIQA	0.36207663	0.48369785

The results enforce the idea that there is a need to develop better suited metrics for solutions that utilize generative networks, being the no-reference metric of relevancy since some computer vision tasks might not allow for the use of a ground truth image by design.

VI. Summary and Future Work

The focus of this Thesis was the development of a subjective quality assessment as well as the identification of the objective quality metrics suitable for GAN-based models. In order to do the subjective assessment, first were selected 3 GAN-based solutions that performed different image processing tasks. These solutions were ESRGAN, a solution that performed super resolution, HiFiC, a compression solution and lastly ArNet an artefact removal solution. These solutions were then applied as seen adequate to fit the requirements established for the assessment.

The subjective assessment provided insight to the performance of each solution as well as the impact of increasing the artefacts produced by the GAN on the perceive image quality. It was clear from the collected data that the participants of this study found the *naturalness* of the GAN-derived images agreeable and, when comparing the GAN compression solution with the HEVC-Intra codec, the subjects displayed an astonishing preference for the HiFiC solution.

Regarding the objective metrics tested during this study and how well they related with the subjective image scores, the statistics demonstrated the poor performance of almost every metric when applied to our subjective data. This was highlighted by the low coefficients of correlation obtained between the objective metrics and the subjective scores. It was also seen, as expected, that for no-reference images the correctness of quality prediction is worst then the one displayed by full-reference metrics.

With the information collected from this study was created a database build for IQA consisting of the generative images and the results of the subjective and objective evaluations. As well as a web application that allows the application of subjective assessments that follow the forced choice methodology.

Following the lack of adequate metrics to evaluate these GAN-based solutions one possible avenue for future work could consist of exploring other metrics already in literature or the attempt to develop a new metric that would be better fit for this task.

REFERENCES

[1] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville and Y. Bengio, "Generative Adversarial Networks," arXiv:1406.2661v1 [stat.ML] , June 2014.

[2] E. Agustsson, M. Tschannen, F. Mentzer, R. Timofte and L. V. Gool, "Generative Adversarial Networks for Extreme Learned Image Compression," in *International Conference on Computer Vision*, Seoul, Korea (South), November 2019.

[3] M. Mirza and S. Osindero, "Conditional Generative Adversarial Nets," arXiv:1411.1784v1 [cs.LG], November 2014.

[4] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang and W. Shi, "Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network," in *Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, July 2017.

[5] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," in *Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, June 2016.

[6] L. Galteri, L. Seidenari, M. Bertini and A. D. Bimbo, "Deep Universal Generative Adversarial Compression Artifact Removal," *IEEE Transactions on Multimedia,* vol. 21, no. 8, pp. 2131-2145, August 2019.

[7] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," arXiv:1409.1556v6 [cs.CV], April 2015.

[8] J. Ascenso and P. Akayzi, "JPEG AI Image Coding Common Test Conditions," in *ISO/IEC JTC 1/SC 29/WG 1 N84035, 84th Meeting*, Brussels, Belgium, July 2019.

[9] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao and C. Change Loy, "ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks," in *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 2018.

[10] A. Jolicoeur-Martineau, "The relativistic discriminator: a key element missing from standard GAN," *arXiv preprint arXiv:1807.00734,* 2018.

[11] F. Mentzer, G. D. Toderici, M. Tschannen and E. Agustsson, "High-Fidelity Generative Image Compression," *Advances in Neural Information Processing Systems,* vol. 33, 2020.

[12] T.-Y. Zitnick, L. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár and C. Lawrence, "Microsoft COCO: Common objects in context," in *European Conference on Computer Vision*, Zurich, Switzerland, September 2014.

[13] F. Mameli, M. Bertini, L. Galteri and A. Del Bimbo, "Image and Video Restoration and Compression Artefact Removal Using a NoGAN Approach," in *Proceedings of the 28th ACM International Conference on Multimedia*, New York, NY, USA, 2020.

[14] J. Antic, J. Howard and U. Manor, *Decrappification, DeOldification, and Super Resolution,* 2019.

[15] International Telecommunication Union, "Methodology for the subjective assessment of the quality of television images," *ITU-R BT.500-14*, October 2019.

[16] H. Ko, D. Y. Lee, S. Cho and A. C. Bovik, "Quality Prediction on Deep Generative Images," *IEEE Transactions on Image Processing,* vol. 29, p. 5964–5979, April 2020.

[17] S. Tilkov and S. Vinoski, "Node.js: Using JavaScript to Build High-Performance Network Programs," *IEEE Internet Computing,* vol. 14, pp. 80-83, 2010.

[18] OpenJS Foundation, "Express JS," [Online]. Available: http://expressjs.com/. [Accessed 4 March 2021].

[19] MongoDB, Inc., "MongoDB," [Online]. Available: https://www.mongodb.com/. [Accessed 4 March 2021].

[20] "PUG," [Online]. Available: pugjs.org. [Accessed 9 March 2021].

[21] K.-T. Chen, C.-C. Wu, Y.-C. Chang and C.-L. Lei, "A Crowdsourceable QoE Evaluation Framework for Multimedia Content," in *Proceedings of the 17th ACM International Conference on Multimedia*, New York, NY, USA, 2009.

[22] Z. Zhang, J. Zhau, N. Liu, X. Gu and Y. Zhang, "An improved pairwise comparison scaling method for subjective image quality assessment," in *2017 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB),* 2017.

[23] A. Hagberg, D. Schult and P. Swart, "NetworkX - Network Analysis in Python," [Online]. Available: https://networkx.org/. [Accessed 7 June 2021].

[24] P. Rao and L. Kupper, "Ties in Paired-Comparison Experiments: A Generalization of the Bradley-Terry Model," *Journal of the American Statistical Association,* vol. 62, p. 194–204, 1967. July 2021.

[25] L. Maystre and M. Grossglauser, "ChoiceRank: Identifying Preferences from Node Traffic in Networks," in *Proceedings of the 34th International Conference on Machine Learning,* 2017.

[26] Z. Wang, A. C. Bovik, H. R. Sheikh and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing,* vol. 13, pp. 600-612, 2004.

[27] Z. Wang, E. P. Simoncelli and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *The Thrity-Seventh Asilomar Conference on Signals, Systems Computers, 2003*, 2003.

[28] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Transactions on Image Processing,* vol. 15, pp. 430-444, 2006.

[29] A. Mittal, A. K. Moorthy and A. C. Bovik, "Blind/Referenceless Image Spatial Quality Evaluator," in *2011 Conference Record of the Forty Fifth Asilomar Conference on Signals, Systems and Computers (ASILOMAR)*, 2011.

[30] W. Zhang, K. Ma, G. Zhai and X. Yang, "Uncertainty-Aware Blind Image Quality Assessment in the Laboratory and Wild," *IEEE Transactions on Image Processing,* vol. 30, pp. 3474-3486, 2021.

[31] S. Su, Q. Yan, Y. Zhu, C. Zhang, X. Ge, J. Sun and Y. Zhang, "Blindly Assess Image Quality in the Wild Guided by a Self-Adaptive Hyper Network," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[32] S. V. D. Walt, J. Schönberger, J. Nunez-Iglesias, F. Boulogne, J. D. Warner, N. Yager, E. Gouillart and T. Yu, " scikit-image: Image processing in Python," *PeerJ,* vol. 2, p. 453, 2014.

[33] A. Khalel, "Sewar," [Online]. Available: https://github.com/andrewekhalel/sewar. [Accessed 4 July 2021].

[34] R. Ocampo, "Image Quality," [Online]. Available: https://github.com/ocampor/image-quality. [Accessed 2021 July 7].

[35] V. Batagelj and A. Mrvar, "A subquadratic triad census algorithm for large sparse networks with small maximum degree," *Social Networks,* vol. 23, pp. 237-243, 2001.

[36] L. Maystre, "choix," [Online]. Available: https://github.com/lucasmaystre/choix. [Accessed 4 July 2021].

.