LANBIQUE: LANguage-based Blind Image QUality Evaluation

LEONARDO GALTERI, LORENZO SEIDENARI, PIETRO BONGINI, MARCO BERTINI, ALBERTO DEL BIMBO, MICC - Università degli Studi di Firenze, Italy

Image quality assessment is often performed with deep networks which are fine-tuned to regress a human provided quality score of a given image. Usually, this approaches may lack generalization capabilities and, while being highly precise on similar image distribution, it may yield lower correlation on unseen distortions. In particular they show poor performances whereas images corrupted by noise, blur or compressed have been restored by generative models. As a matter of fact, evaluation of these generative models is often performed providing anecdotal results to the reader. In the case of image enhancement and restoration, reference images are usually available. Nonetheless, using signal based metrics often leads to counterintuitive results: highly natural crisp images may obtain worse scores than blurry ones. On the other hand, blind reference image assessment may rank images reconstructed with GANs higher than the original undistorted images. To avoid time consuming human based image assessment, semantic computer vision tasks may be exploited instead.

In this paper we advocate the use of language generation tasks to evaluate the quality of restored images. We refer to our assessment approach as LANguage-based Blind Image QUality Evaluation (LANBIQUE). We show experimentally that image captioning, used as a downstream task, may serve as a method to score image quality, independently of the distortion process that affects the data. Captioning scores are better aligned with human rankings with respect to classic signal based or No-Reference image quality metrics. We show insights on how the corruption, by artifacts, of local image structure may steer image captions in the wrong direction.

CCS Concepts: • **Computing methodologies → Computer vision tasks**; **Scene understanding**; **Image representations**; *Object recognition*; Image compression.

Additional Key Words and Phrases: image quality enhancement, image captioning, image quality evaluation, GAN, generative models evaluation

1 INTRODUCTION

In the last years, models able to generate novel images by implicit sampling from the data distribution have been proposed [16]. While these models are extremely appealing, generating for example photo realistic faces [22] or landscapes [35], they are hard to be evaluated. Often anecdotal qualitative examples are presented to the reader with little quantitative and objective evidence, and evaluation of generative models is still undergoing a debate regarding how to perform it. The idea of using a computer vision classifier to evaluate the veracity of a generated images was first proposed in [39]. The authors propose the Inception Score (IS), which is obtained applying the Inception model [42] to every generated image in order to obtain the conditional label distribution $p(y|x)$. Realistic images should contain one or few well defined objects therefore leading to a low entropy in the conditional label distribution $p(y|x)$. An improved evaluation metric, named Frechét Inception Distance (FID) has been proposed by [18]. The authors show that FID is more consistent than Inception Score with increasing

---

Author's address: Leonardo Galteri, Lorenzo Seidenari, Pietro Bongini, Marco Bertini, Alberto Del Bimbo, MICC - Università degli Studi di Firenze, Firenze, Italy, [name.surname]@unifi.it.

1551-6857/2022/5-ART $15.00
https://doi.org/10.1145/3538649

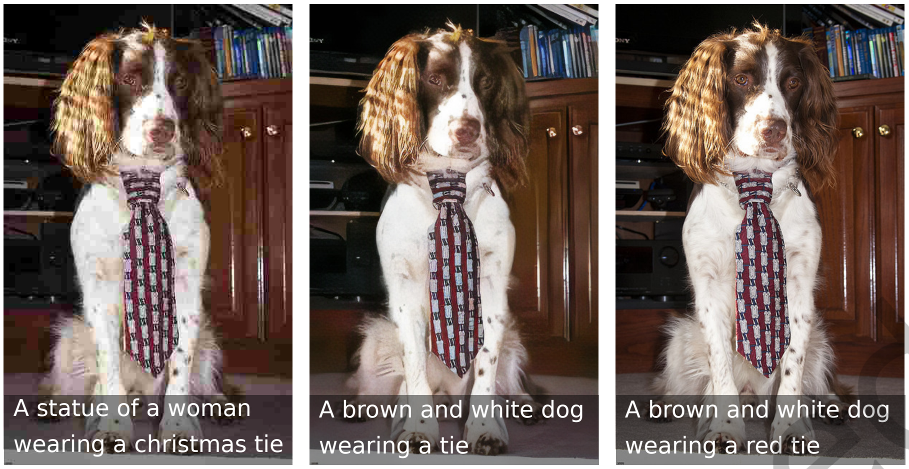

A statue of a woman wearing a christmas tie
A brown and white dog wearing a tie
A brown and white dog wearing a red tie

Fig. 1. Caption generated on Compressed, Reconstructed and Original image (left to right) using [2]. Sample ground truth caption: “A brown and white dog wearing a necktie”. Best viewed in color on computer screen.

disturbances and human judgment. FID performs better as an evaluation metric since it also exploits the statistics of the real images.

Recently [6, 25, 41] have specifically addressed methods to evaluate Generative Adversarial Networks (GANs). In [41] have been proposed two methods that evaluate diversity and quality of generated images using classifiers trained and tested on generated images. In [5] the authors trained an Auxiliary Classifier GAN to generate new distorted samples to train a shallow quality evaluator to solve the lack of data in the standard datasets. In [6] a discussion of 24 quantitative and 5 qualitative measures for evaluating generative models is provided, including IS and FID, image retrieval and classification performance.

Apart from generating new images, GANs can be effectively used to enhance the visual quality of images that have been subjected to some degradation, like noise or compression. In this use case the generator network is conditioned with the degraded input, and it produces an enhanced version. In [25] it is observed that many existing image quality assessment (IQA) algorithms do not correctly assess GAN generated content, especially when considering textured regions; this is due to the fact that although GANs generate very realistic images that may look like the original one, they match them poorly when considering pixel-based metrics. The proposed metric, called SSQP (Structural and Statistical Quality Predictor), is based on the “naturalness" of the image.

Subjective metrics, such as Mean Opinion Score are obtained by presenting images to several human evaluators and asking for a subjective score on the image quality. Such mean of measuring image quality is possibly the best choice but has the obvious drawback of human annotators need and the related cost in terms of time and money to rank a high volume of data.

Regarding the evaluation of image enhancement methods, only recently semantic computer vision tasks have been proposed for image quality assessment. The motivation behind this choice is twofold. On the one hand, since images are often processed by algorithms, it is intrinsically interesting to evaluate the performance of such algorithms on degraded and restored images; to this regard, it has to be noted that MPEG leads an activity on Video Coding for Machines (VCM), that aims to standardize video codecs in the case where videos are consumed by algorithms. On the other hand, we assume that semantic computer vision tasks lead to a more robust evaluation protocol. In previous works object detection and segmentation have been used to assess image enhancement [13, 14, 51].

In this paper, we introduce a novel image quality assessment method based on language models. To the best of our knowledge, language has never been used to evaluate the quality of images. We refer to the new approach as LANguage-based Blind Image QUality Evaluation (LANBIQUE). Fig. 1 shows the gist of the proposed approach: the effects of image compression lead to a wrong captioning of the image on the left with respect to the original high quality image on the right; captioning an image that has been obtained enhancing the compressed image with a GAN-based approach (center) leads to a caption that is very similar to the caption of the high quality image. The main contributions of our work are the following:

- LANBIQUE show consistency across different captioning algorithms [2, 11] and language similarity metrics. Interestingly, improving the language generation model also improves the correlation between our score and MOS.
- Experiments shows that LANBIQUE does not suffer from drawbacks of common Full-Reference and No-Reference metrics when evaluating GAN enhanced images and keeps a high accordance with human scores for compressed and for images restored via deep learning.

In this extended version, we propose the following improvement with respect to [15].

- We show that LANBIQUE can be used also for distortions different from JPEG compression.
- We tested LANBIQUE on the larger and more diverse PieAPP dataset, showing strong results against learning and non-learning based methods.
- Finally, the basic version of LANBIQUE is extended in order to make it possible to work also without a reference image. To get to this goal we employ a blind restoration GAN, which can restore images without the knowledge nor the intensity of the distortion, to recover a pseudo-reference image.

The rest of the paper is organized as follows: in Section 2 we describe the related works. In Section 3 we briefly discuss about prior GAN-based image restoration approaches. In Section 4 we describe LANBIQUE in detail. In Section 5 we show experimental results of LANBIQUE on different settings and datasets. Finally, in Section 6 we draw the conclusions about our approach.

2 RELATED WORK

Full-Reference quality assessment. When dealing with image restoration tasks, a reference image is often available to perform evaluation. Full-Reference image quality assessment is an evaluation protocol which uses a reference version of an image to compute a similarity. Popular metrics are Peak Signal-to-Noise Ratio (PSNR) and Mean Squared Error (MSE). However, these metrics have been often criticized because they are not consistent with human perceived quality of images [49]. SSIM, a metric of structural similarity, has been proposed to overcome this limitation. Unfortunately, as will be shown in the following, even SSIM is too simplistic to capture human perceived quality of images; moreover, distortion metrics have been shown to be at odds with high perceptual quality. Blau and Michaeli [4] propose a generalization of rate-distortion theory which takes perceptual quality into account, and study the three-way trade-off between rate, distortion, and perception. The authors show that aiming at obtaining a high perceptual quality leads to an elevation of the rate-distortion curve and thus requires to make a sacrifice in either the distortion or the rate of the algorithm.

No-Reference quality assessment. No-Reference image assessment techniques are devised in the realistic scenario in which image quality must be estimated without accessing an original high quality or uncompressed version of the image itself. Recent No-Reference image quality assessment methods are based on natural scene statistics (NSS), computed in the spatial domain. Instead of extracting distortion specific statistics such as the amount of blur or ringing in an image, they look at the statistics of locally normalized luminance in order to estimate the loss in image naturalness. These metrics are designed and optimized in order to be highly correlated with human subjective metrics. Pei and Cheng [36] train a random forest for IQA using the features extracted from

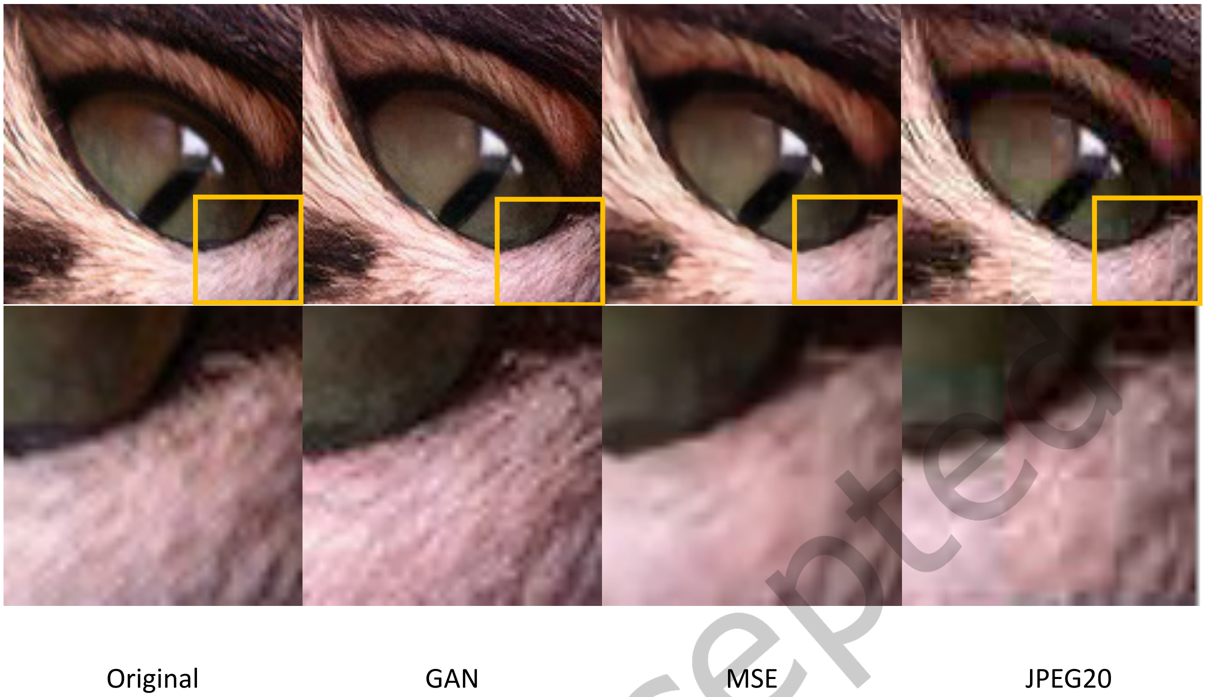


Fig. 2. Qualitative comparison of reconstruction methods: GAN produces images more pleasant for the human eye. Best viewed in color and zoomed on computer screen. GAN: GAN-based restoration using perceptual loss. MSE: CNN-based restoration using MSE loss; JPEG 20: JPEG compression with quality factor 20;

the difference of Gaussian (DOG) bands and demonstrate it highly correlates with human visual system. Lukin *et al.* [29] fuse the outcome of several quality assessment systems by training a neural network. Kim and Lee [24] propose a Full-Reference framework that aims to learn the human visual sensitivity by leveraging distorted images, objective error maps and subjective scores. Bosse *et al.* [7] propose a learned approach for image quality assessment that incorporates an optional joint optimization of weighted average patch aggregation implementing a method for pooling local patch qualities to global image quality. In [28] Liu *et al.* address the problem of the lack of data in the standard IQA datasets with a siamese network that learns from rankings. This approach obtains impressive results. In the last few years with the advent of new large datasets [19, 37] for Image Quality Assessment, No-Reference and Full-Reference transformer-based approaches were deployed obtaining very high performances [10, 52].

3 IMAGE RESTORATION

Even if this work does not propose novel image restoration approaches, to make the paper self-contained here we formalize the image restoration or enhancement task. The main motivation that lead us to work on an alternative to image quality assessment is the poor performance of standard IQA methods on images that have been enhanced by GANs, e.g. for denoising [23, 44], deblurring [43, 53] or compression artefact removal [14, 30, 45]. Furthermore, we leverage image restoration as a tool to extend the capabilities of LANBIQUE in order to evaluate those images that lack an uncorrupted high quality counterpart, extending our approach to the No-Reference scenario, as show in Sect.4.3.

Problem formulation. Given some image processing algorithm D, such as JPEG image compression, a distorted image is defined as $I_{LQ} = D(I_{HQ})$, where I_{HQ} is a high quality image undergoing the distortion process, image enhancement aims at finding a restored version of the image $I_R \approx G(I_{LQ})$. In this work we use two image enhancement networks, one that is specific for JPEG artifacts [14], and a more generic approach, which can work without prior knowledge of the degradation [47].

In [14] Galteri *et al.* try to learn a generative model G which, conditioned on the input distorted images, is optimized to invert the distortion process D so that $G \approx D^{-1}$. Their generator architecture is loosely inspired by [17]. They employ LeakyReLU activations and 15 residual layers in a fully convolutional network. The final image is obtained by a nearest neighbor upsampling of a convolutional feature map and a following stride-one convolutional layer to avoid grid-like patterns possibly stemming from transposed convolutions.

The set of weights ψ of the D network are learned by minimizing:

$$\mathcal{L}_d = -\log\left(D_\psi\left(I_{HQ}|I_{LQ}\right)\right) - \log\left(1 - D_\psi\left(I_R|I_{LQ}\right)\right)$$

where I_{HQ} is the uncompressed or high-quality image, I_R is the restored image created by the generator and I_{LQ} is a compressed image.

The generator is trained combining a perceptual loss with the adversarial loss:

$$\mathcal{L}_{AR} = \mathcal{L}_P + \lambda\mathcal{L}_{adv}. \tag{1}$$

where $\mathcal{L}_{adv}$ is the standard adversarial loss:

$$\mathcal{L}_{adv} = -\log\left(D_\psi\left(I_R|I_{LQ}\right)\right) \tag{2}$$

that rewards solutions that are able to mislead the discriminator, and $\mathcal{L}_p$ is a perceptual loss based on the distance between images computed projecting I_{HQ} and I_R on a feature space by some differentiable function ϕ and taking the Euclidean distance between the two feature representations:

$$\mathcal{L}_P = \frac{1}{W_f H_f}\sum_{x=1}^{W_f}\sum_{y=1}^{H_f}\left(\phi\left(I_{HQ}\right)_{x,y} - \phi\left(I_R\right)_{x,y}\right)^2 \tag{3}$$

They employ a generator inspired by [17], with a residual architecture using LeakyReLU activations, BatchNormalization [20] and Nearest-neighbour upsampling layer is used to recover original size [33], and a fully convolutional Discriminator. In [14] it has been shown that using a GAN approach instead of direct training of the network for image enhancement, results in improved subjective perceptual similarity to original images and, more importantly, in much improved object detection performance. Qualitative examples of GAN and direct training method are shown in Fig. 2.

Real-ESRGAN [47] is a more recent approach, that has the advantage of not requiring to know the type of distortion nor the intensity of it in advance to restore an image. In [47] Wang *et al.* introduce a high-order degradation modeling process to better simulate complex real-world degradations. Differently from [14] they use a U-Net discriminator with spectral normalization to increase discriminator capability and stabilize the training dynamics. As in ESRGAN [48] the generator is built by several residual-in-residual dense blocks (RRDB).

4 EVALUATION PROTOCOL

Classic Full-Reference image quality evaluation methods rely on the similarity between an image which has been processed by some algorithm D and a reference undistorted image. Considering the use case of image enhancement of an image that was compressed, GANs are a good solution since they are great at filling in high frequency realistic details in image enhancement tasks; in this case the resulting enhanced image is compared to

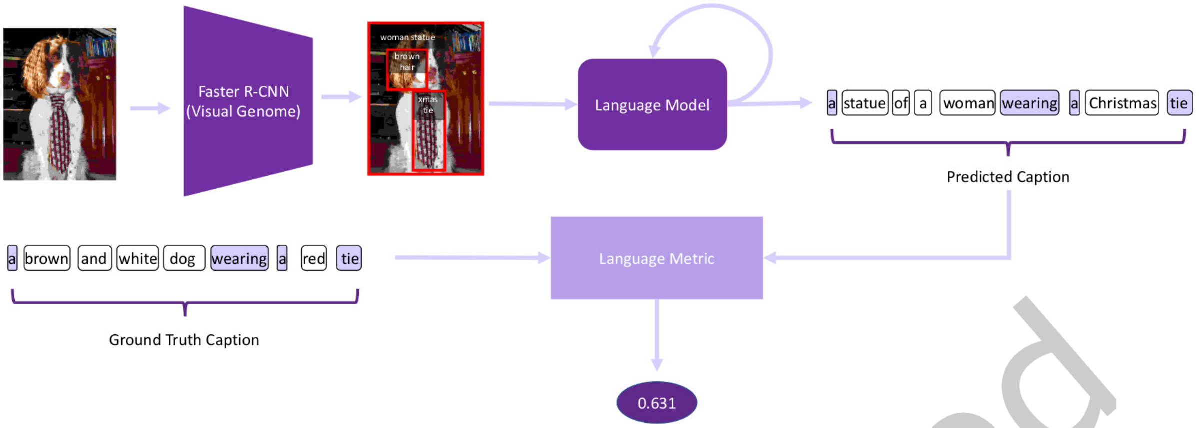


Fig. 3. Overview of LANBIQUE. An image is first processed by an object detector, each box feature is then fed to a captioning model[2, 11]; then a metric for captioning evaluation is used to score the quality of the image. In this example a highly JPEG corrupted images yields a *low* CIDEr score of 0.631.

the reference. Unfortunately, when using classical MSE based Full-Reference metrics such as SSIM and PSNR GAN restored images yield lower performance as can be seen in Tab. 2, although they appear as "natural" and pleasant to human evaluators, as also seen in examples of Fig.2. For this reason, in [13, 14] semantic tasks are used to evaluate the quality of restored images. Measuring the performance of a semantic task such as detection on restored images gives us an understanding of the "correctness" of output images. Given some semantic task (e.g. object detection), a corresponding evaluation metric (e.g. mAP) and a dataset, the evaluation protocol consists in measuring the variation of such metric on different versions of the original image. Interestingly, this evaluation methodology gives hints on what details are better recovered by GANs.

In certain cases, detection is a task describing scene semantics in a very approximate fashion; usually detectors do not degrade for object classes that are clearly identifiable by their shape since even high distortions in the image are not able to hide such features. The gain in image quality provided by GANs, according to object detection based evaluation, resides in producing high quality textures for deformable objects (e.g. cats, dogs, etc).

In this paper we advocate the use of a language generation task for evaluating image enhancement. The idea is that captioning maps the semantics of images into a much finer and rich label space represented by short sentences. To be able to obtain a correct caption from an image many details must be identifiable.

4.1 Evaluation with Reference Captions

We devise the following evaluation protocol for image enhancement. We pick an image captioning algorithm $\mathcal{A}$. Image captioning is the task of generating a sequence of words, possibly grammatically and semantically correct, describing the image in detail. Given a set of reference captions S and the caption generated from an input image $\mathcal{A}(I)$, we want to measure their similarity with a language metric $\mathcal{D}$:

$$\text{LANBIQUE}(\mathcal{D}, \mathcal{A}; I, S) = \mathcal{D}(\mathcal{A}(I), S) \tag{4}$$

We look at the performance of a captioning algorithm $\mathcal{A}$ on different versions of a dataset (e.g. COCO): compressed, original and restored. The pipeline of this evaluation approach is depicted in Fig. 3.

In particular, we analyze results from two highly performing captioning methods [2, 11] which combine a bottom-up model of visual entities and their attributes in the scene with a language decoding pipeline. Both methods are trained over several steps incorporating semantic knowledge at different levels of granularity. In particular, the bottom-up region generator is based on Faster R-CNN [38] which is based on a feature extractor pre-trained on ImageNet [12] and then fine-tuned to predict object entities and their attributes using the Visual Genome dataset [26]. In [2], further knowledge is incorporated into the model by training the caption generation model using a first LSTM as a top-down visual attention model and a second level LSTM as a language model. Meshed memory transformers [11] share the exact same visual backbone as [2] but exploit a stack of memory-augmented visual encoding layers and a stack of decoding layers to generate caption tokens.

No matter how captioning models are optimized, our results show that the behavior of the captioning model for image quality assessment is consistent over several metrics as shown in Tab. 1.

Captioning is evaluated with several specialized metrics measuring the word-by-word overlap between a generated sentence and the ground truth [34], in certain cases including the ordering of words [3], considering n-grams and not just words [27, 46] and the semantic propositional content (SPICE [1]). These metrics evaluate the similarity with respect to a set of reference captions S, that is usually composed of five references.

4.2 Evaluation without Reference Captions

Unfortunately, in most of the cases reference captions are not available as they often must be collected with great expense of effort and resources; in fact, standard datasets used for image quality evaluation do not include captions. However, it is possible to evaluate any kind of test image with our language based approach by modifying the pipeline. The idea is that the reference image is enough high quality to provide a valid caption for the evaluation of LANBIQUE. We caption the reference image I_{HQ} using the same captioner $\mathcal{A}$ we use for the test image I, then we maintain the same procedure we previously described:

$$\text{LANBIQUE-NC}(\mathcal{D}, \mathcal{A}; I, I_{HQ}) = \mathcal{D}(\mathcal{A}(I), \mathcal{A}(I_{HQ})) \tag{5}$$

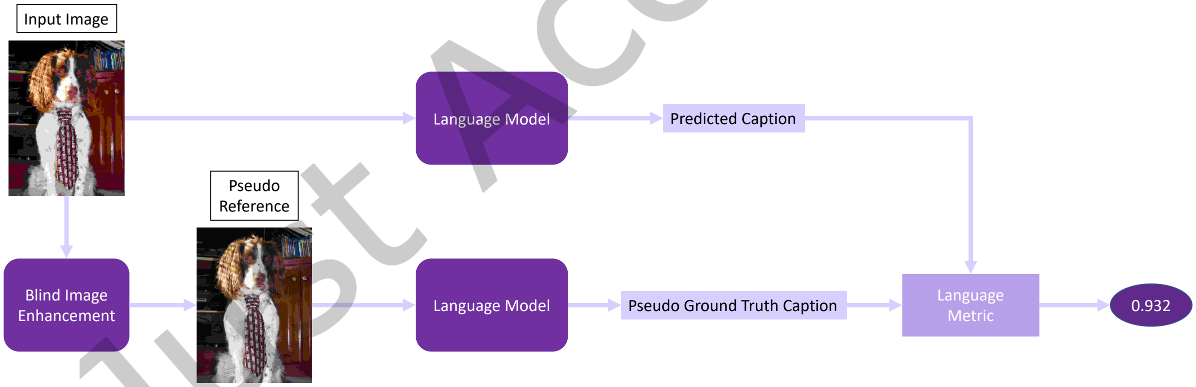


Fig. 4. LANBIQUE without a reference caption available. The reference image is captioned as well by the same language model to generate a description of the image. This output is used as pseudo ground truth caption and compared to the predicted caption.

This evaluation approach is represented in Fig. 4. Since we change the evaluation pipeline with respect to the previous case, we argue that there may be a drawback with respect to the original version of the approach. As a matter of fact, modern captioners provide just one description per image and this means that the computation of $\mathcal{D}$ metric is done just between two sentences. However, this does not affect the performance of our approach significantly, provided that the $\mathcal{A}$ generates high quality captions.

4.3 No-Reference Evaluation

In this section we show how our approach can be extended to work in a No-Reference setting. In many occasions we may not have a high quality image available to be compared with the one to be tested. For this reason, we modify our language based pipeline by adding an additional blind restoration module $\mathcal{R}$. We assume that the images to be tested are corrupted by one or a combination of unknown distortions that are responsible of a global reduction of the visual quality. In this extended model, our aim is to restore corrupted input image I in order to use the enhanced version as the reference image. After this operation is completed, we are able to feed both the corrupted image and the restored one to the same captioning module, hence we generate their text descriptions and finally we calculate the ultimate score based on some language metric $\mathcal{D}$:

$$\text{LANBIQUE-NR}(\mathcal{D}, \mathcal{A}, \mathcal{R}; I) = \mathcal{D}(\mathcal{A}(I), \mathcal{A}(\mathcal{R}(I))) \tag{6}$$

This No-Reference approach is depicted in Fig.5

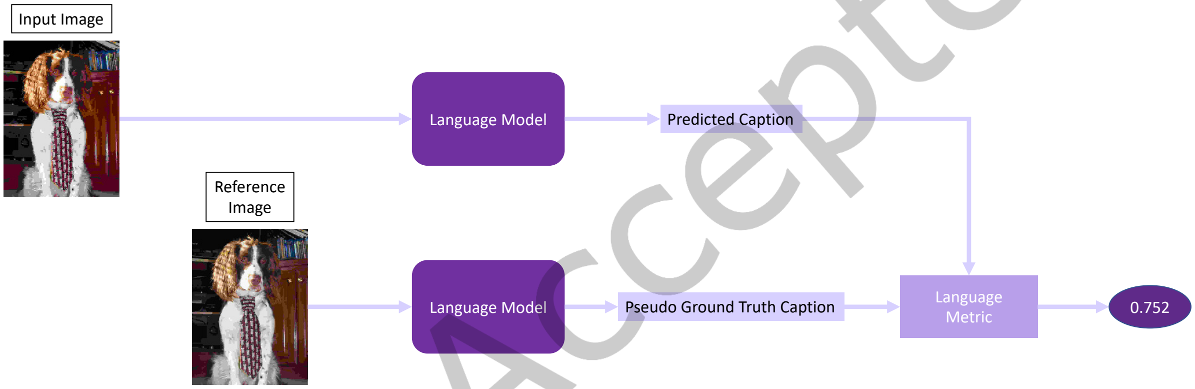


Fig. 5. LANBIQUE in the No-Reference setting. A blind image enhancement method is used to recover a high quality version of the image, then a captioning model is applied to both images. Input image predicted caption is then compared with the pseudo ground truth caption obtained from the restored image.

Typically, image distortions are not known a priori so it may be a difficult task to train many networks capable of handling all the possible combinations of corruption processes and then select the best one for a specific restoration. For this reason, we choose to train a single network following a degradation model, so that it can restore most types of distorted images and recover their original quality as best as possible. In order to ensure a good output quality, we employed Real-ESRGAN [47] as the restoration module. We have modified the original model by adding JPEG2000 in the training procedure, then we have fine-tuned a pre-trained version of such network with the new introduced distortion.

In most of the cases, recovered images represent a solid reference for our evaluation model, as they are very close to real images from the point of view of human perception. In this setup, our LANBIQUE-NR assigns high scores to slightly distorted images, as their reconstruction is likely very perceptually close, and the captions generated are pretty close. On the other hand, highly distorted images are transformed into better quality data that differ significantly from input. In this case, the captions between the two versions may differ much more, thus leading to lower scores of language metrics.

4.4 Subjective evaluation

In this evaluation we assess how images obtained with the selected GAN based restoration method [14] are perceived by a human viewer, evaluating in particular the preservation of details and overall quality of an image. In total, 16 viewers have participated to the test, a number that is considered enough for subjective image quality evaluation tests [50]; no viewer was familiar with image quality evaluation or the approaches proposed in this work. A Single-Stimulus Absolute Category Rating (ACR) experimental setup has been developed using avrateNG[1], a tool designed to perform subjective image and video quality evaluations. We asked participants to evaluate images' quality using the standard 5-values ACR scale (1=bad, up to 5=excellent). A set of 20 images is chosen from the COCO dataset, selecting for each image three versions: the original image, a JPEG compressed version with QF=10 (high compression quality factor) and the restored version of the JPEG compressed image with QF=10 compressed image; this results in a set of 60 images. Each image was shown for 5 seconds, preceded and followed by a grey image, also shown for 5 seconds. Considering our estimation of test completion time, we chose this amount of images to keep each session under 30 minutes as recommended by ITU-R BT.500-13 [21].

To select this small sample of 20 images to be as representative as possible of the whole dataset $\mathcal{D}$ for the captioning performance we operate the following procedure. Let $\mu^*(v)$ and $\sigma^{2^*}(v)$ be the mean and variance of a captioning metric score (in this case we used CIDEr) for a given version v of the image i. We iteratively extract 20 random image ids, yielding set $\mathcal{D}^*$ out of the whole 5,000 testing set from the Karpathy split, without repetition. We attempt to minimize

$$e_\mu = \frac{1}{|\mathcal{D}^*|}\sum_{i\in\mathcal{D}^*}\sum_{v\in\mathcal{V}_i}|\mu^*(v) - \overline{\mu}| \tag{7}$$

and

$$e_{\sigma^2} = \frac{1}{|\mathcal{D}^*|}\sum_{i\in\mathcal{D}^*}\sum_{v\in\mathcal{V}_i}|\sigma^{2^*}(v) - \overline{\sigma}^2| \tag{8}$$

by iterative resampling images until we find e_μ and e_{σ^2} such that $e_\mu \leq 10^{-3}$ and $e_{\sigma^2} \leq 10^{-4}$. $\mathcal{V}_i$ is the set of different versions of an image i in the smaller dataset $\mathcal{D}^*$, namely: JPEG compressed at QF=10 (referred to as JPEG 10 in the following), its GAN reconstruction and the original uncompressed image; and $\overline{\mu}$ and $\overline{\sigma}^2$ are the mean and variance of the considered captioning metric computed on the whole set of images $\mathcal{D}$. The selected images contain different subjects, such as people, animals, man-made objects, nature scenes, etc. Both the order of presentation of the tests for each viewer, and the order of appearance of the images were randomized.

5 EXPERIMENTAL RESULTS

5.1 Results on JPEG Artefacts

First, we study in detail the behavior of LANBIQUE on a single distortion. This way we can easily control the amount of image corruption and evaluate the behavior of our metric on GAN restored images.

Results with reference captions. In order to use a dataset of images with a set of associated captions, we selected the 5,000 images testing set from the Karpathy split of COCO dataset [9]. The images have then been compressed at different JPEG Quality Factors (QF), and then they have been reconstructed using the GAN approach of [14]. In Tab. 1 we report results of LANBIQUE using various captioning metrics $\mathcal{D}$. Interestingly, all metrics show that captions over reconstructed images (REC rows) are better with respect to caption computed over compressed images (JPEG rows). This shows that image details that are compromised by the strong compression induce errors in the captioning algorithm. On the other hand, the GAN approach is able to recover an image which is not

---

[1] https://github.com/Telecommunication-Telemedia-Assessment/avrateNG

only pleasant to the human eye but recovers details which are also relevant to a semantic algorithm. In Fig. 1 we show the difference of captions generated by [2] over original, compressed and restored images. A human may likely succeed in producing an almost correct caption for highly compressed images, nonetheless state-of-the art algorithms are likely to make extreme mistakes which are instead not present on reconstructed images.

Table 1. Evaluation of image restoration over compression artifacts with GAN using LANBIQUE with different captioning metrics (best results highlighted in bold). For each metric we denote higher(↑) or lower(↓) is better. JPEG q indicates a JPEG compressed image with $QF = q$ (e.g. 10), while (REC q) indicates the corresponding reconstruction using [14]. Captions created from reconstructed images obtain a better score for every metric.

QUALITY	BLEU_1↑	METEOR↑	ROUGE↑	CIDEr↑	SPICE↑
JPEG 10	0.589	0.173	0.427	0.496	0.103
REC 10	**0.730**	**0.253**	**0.527**	**1.032**	**0.189**
JPEG 20	0.709	0.241	0.513	0.937	0.174
REC 20	**0.751**	**0.266**	**0.543**	**1.105**	**0.201**
JPEG 30	0.740	0.258	0.535	1.054	0.194
REC 30	**0.757**	**0.269**	**0.549**	**1.133**	**0.205**
JPEG 40	0.748	0.263	0.542	1.087	0.200
REC 40	**0.758**	**0.270**	**0.549**	**1.132**	**0.206**
JPEG 60	0.755	0.267	0.546	1.117	0.204
REC 60	**0.760**	**0.270**	**0.550**	**1.137**	**0.207**
ORIGINAL	0.766	0.274	0.556	1.166	0.211

In Fig. 6 we show the different performance of captioning algorithms in terms of CIDEr measure on the same split of test of compressed and restored images, considering different quality factors of JPEG. The captioner proposed in [11] outperforms [2] as expected, but interestingly we may observe that the range of CIDEr values of [11] is significantly higher than [2]. We argue that this could be considered a strong feature of our evaluation approach, as a wider range of value may imply that a good captioner is able to predict the image quality in a finer manner than other weaker captioning algorithms.

Fig. 7 shows the bottom-up captioning process performed on an image used in the subjective evaluation. The left image shows the JPEG 10 version, while the right one shows the GAN reconstruction. The images show the bounding boxes of the detected elements. In the first case the wrong detections of indoor elements like “floor” and “wall” are likely reasons for the wrong caption, as opposed to the correct recognition of a “white wave” and “blue water” in the GAN-reconstructed image.

Results without reference captions. A common setting that is used to evaluate image enhancement algorithms is Full-Reference image quality assessment, where several image similarity metrics are used to measure how much a restored version differs with respect to the uncorrupted original image. This kind of metrics, measuring pixel-wise value differences are likely to favor MSE optimized networks which are usually prone to obtain blurry and lowly detailed images.

In certain cases, it is not possible to use Full-Reference quality metrics, e.g. if there’s no available original image. These kind of metrics typically evaluate the “naturalness" of the image being analyzed. In the same setup we used previously, we perform experiments using NIQE and BRISQUE which are two popular No-Reference metrics for images. Interestingly, these metrics tend to favor GAN restored images instead of the original uncompressed ones.

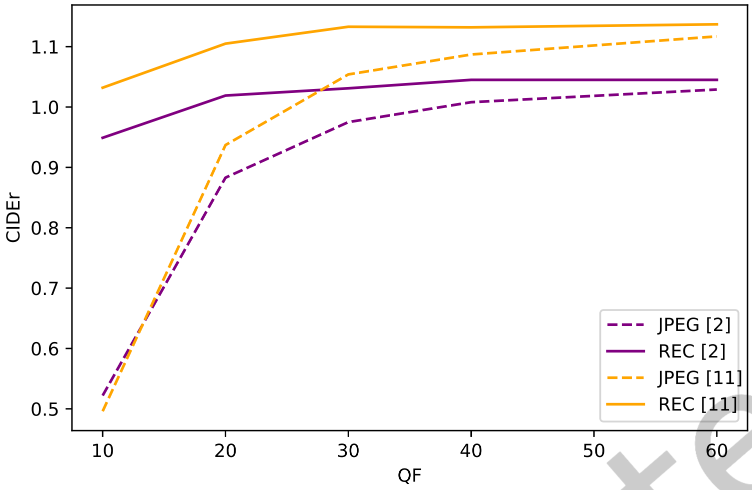
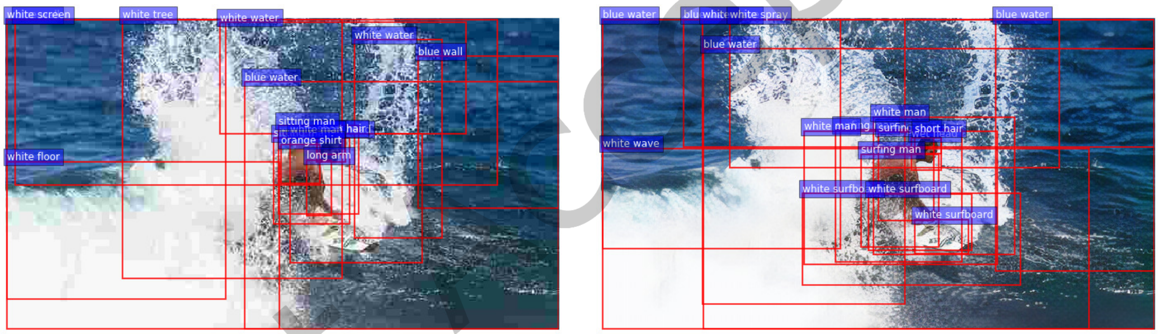


Fig. 6. CIDEr scores using [2] (purple) and [11] (yellow) on compressed and restored images for different QFs from MS-COCO.



A couple of people sitting next to a Christmas tree.

A man riding a wave on a surfboard in the ocean.

Fig. 7. Bottom-Up detection process of captioning on two images: left) JPEG compressed; right) GAN reconstruction. Note that several mistaken detections on the left image are avoided in the right one. In particular on the left "surfboard" is missed and "white floor" and "blue wall" are wrongly detected. These two indoor details are the one that likely mislead the captioning.

Most surprisingly, NIQE and BRISQUE obtain better results when we reconstruct the most degraded version of images (QF 10-20), but these values increase as we reconstruct less degraded images. We believe that BRISQUE and NIQE favor crisper images with high frequency patterns which are distinctive of GAN based image enhancement and they are typically stronger when reconstructing heavily distorted images.

In Tab. 2 we report results on COCO for Full-Reference and No-Reference indexes. In this setup, we compress the original images at different QFs and then we restore them with a QF specific artifact removal GAN. We use the uncompressed image generated caption as ground truth, as in Tab. 3. The results show that, for restored images, PSNR accounts for a slight improvement while SSIM indexes lower than the compressed counterparts.

Table 2. Evaluation using No-Reference and Full-Reference metrics on MS-COCO. For each metric we denote higher(↑) or lower(↓) is better. JPEG q indicates a JPEG compressed image with $QF = q$ (e.g. 10), while (REC q) indicates the corresponding reconstruction using [14]. NIQE and BRISQUE rate better GAN images than the ORIGINAL. SSIM always rate restored images worse than compressed. PSNR shows negligible improvement. [11] and CIDEr have been used by LANBIQUE-NC respectively as language model and language metric.

QUALITY	NIQE↓	BRISQUE↓	PSNR ↑	SSIM↑	LPIPS↓	LANBIQUE-NC ↑
JPEG 10	6.689	52.67	25.45	0.721	0.305	0.542
REC 10	3.488	17.93	25.70	0.718	0.144	1.118
JPEG 20	5.183	43.99	27.46	0.796	0.187	0.956
REC 20	3.884	17.85	27.60	0.784	0.085	1.289
JPEG 30	4.474	37.72	28.61	0.831	0.134	1.165
REC 30	3.601	18.32	28.81	0.819	0.060	1.370
JPEG 40	4.011	33.61	29.41	0.852	0.105	1.260
REC 40	3.680	18.68	29.44	0.836	0.048	1.424
JPEG 60	3.588	28.15	30.71	0.880	0.067	1.366
REC 60	3.885	19.45	30.61	0.862	0.032	1.482
ORIGINAL	3.656	21.79	-	-	-	-

This is an expected outcome, as in [14] it is shown that state of the art results on PSNR can be obtained only when MSE is optimized and on SSIM if the metric is optimized directly. Nonetheless, as can be seen in Fig. 2, GAN enhanced images are more pleasant to the human eye, therefore we should not rely just on PSNR and SSIM for GAN restored images. LANBIQUE, using [11], is in line with LPIPS [54]. Unfortunately, LPIPS, as shown in Tab. 3 has low correlation with scores determined by human perceived quality.

Correlation with Mean Opinion Score. In Fig. 8 *left)* are reported subjective evaluation results as Mean Opinion Scores (MOS) as box plots, showing the quartiles of the scores (box), while the whiskers show the rest of the distribution. The plots are made for the original images, the images compressed with JPEG using a QF=10, and the images restored with the GAN-based approach [14] from the heavily compressed JPEG images. The figure shows that the GAN-based network is able to produce images that are perceptually of much higher quality than the images from which they are originated; the average MOS score for JPEG images is 1.15, for the GAN-based approach is 2.56 and for the original images it is 3.59. The relatively low MOS scores obtained also by the original images are related to the fact that COCO images have a visual quality that is much lower than that of dataset designed for image quality evaluation. To give better insight on the distribution of MOS scores, Fig. 8 *right)* shows the histograms of the MOS scores for the three types of images: orange histogram for the original images, green for the JPEG compressed images and blue for the restored images.

We further show that our language based approach correlates with perceived quality using a IQA benchmark test on the LIVE dataset [40] that consists of 29 high resolution images compressed at different JPEG qualities for a total of 204 images. For each LIVE image a set of user scores is provided indicating the perceived quality of the image. However, no caption is provided in this dataset. For this reason, we consider the output sentences of captioning approaches over the undistorted image as the ground truth in order to calculate the language similarity measures, following the LANBIQUE-NC protocol presented in Sect. 4.2. In Tab. 3 we show the Pearson correlation score of different captioning metrics and other common Full-Reference quality assessment approaches. The experiment shows an interesting behaviour of our approach in terms of correlation. In the first place, we can observe that each captioning metric has a correlation index that is higher or at least comparable with the other Full-Reference metrics. In particular, METEOR and CIDEr perform better than the other metrics independently of

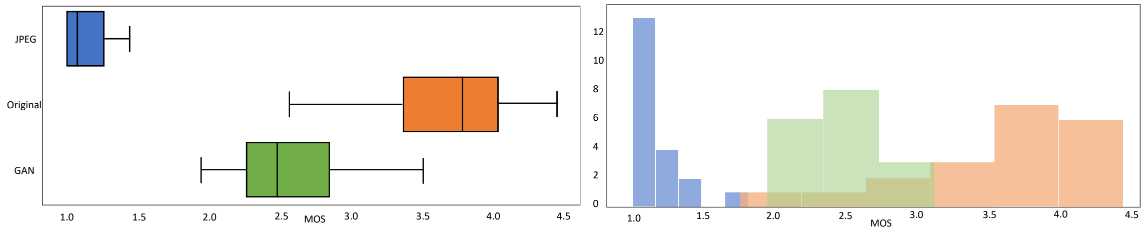


Fig. 8. *Left)* Subjective image quality evaluation of original COCO images (orange), heavily compressed JPEG images (blue) and their restored version obtained with the GAN-based approach (green). Restored images are perceived as having a better quality than their compressed versions. *Right)* Histograms of MOS scores of the three types of images.

Table 3. Pearson score, correlating scores with users' MOS for different captioning metrics and image based Full-Reference approaches on LIVE dataset. CIDEr obtains a superior score with respect to image based methods.

Metric	**LANBIQUE-NC** w/ [11]	**LANBIQUE-NC** w/ [2]
BLEU 1	0.873	0.838
METEOR	0.900	0.846
SPICE	0.895	0.844
ROUGE	0.861	0.832
CIDEr	**0.901**	0.854
PSNR	0.857	
SSIM	0.893	
LPIPS	0.859	

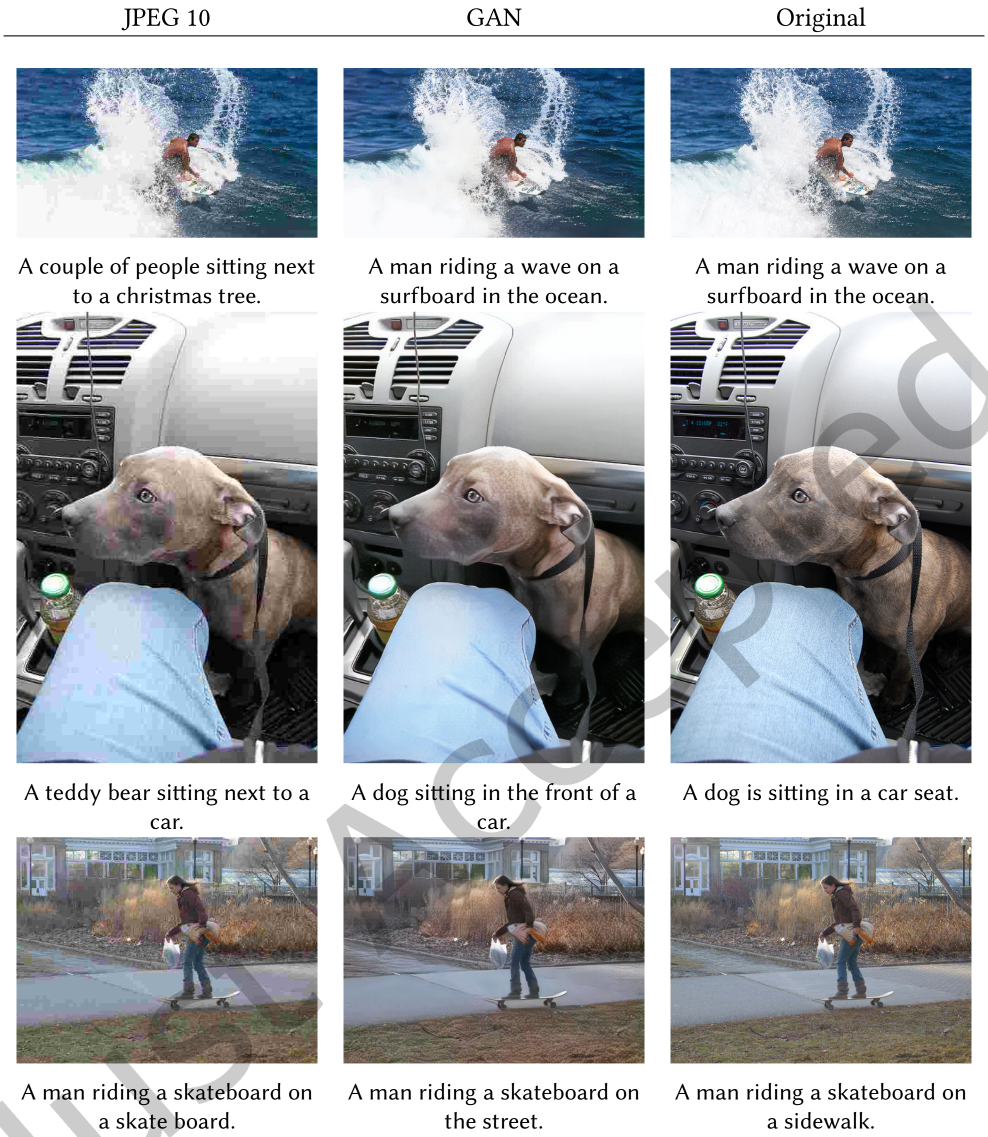
which captioning algorithm is used. In the following experiments LANBIQUE, LANBIQUE-NC and LANBIQUE-NR have been computed using CIDEr metric. Moreover, we observe that the correlation metric significantly improves if we employ a more performing captioner. In this case, the visual features used by the two captioning techniques are exactly the same, the main difference lies in the overall language generation pipeline of the approaches. Hence, we argue that language is effectively useful for quality assessment, and the more a captioning algorithm is capable of providing detailed and meaningful captions the better we could use the generated sentences to formulate good predictions about the quality of images.

In order to better understand what metric could be used instead of human evaluation, we computed the correlation coefficient

$$\rho = \frac{\sum_{i \in \mathcal{D}} (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i \in \mathcal{D}} (x_i - \overline{x})^2 \sum_{i \in \mathcal{D}} (y_i - \overline{y})^2}} \tag{9}$$

between BRISQUE [31], NIQE [32], the proposed LANBIQUE and MOS for all versions of the images. As shown in Tab. 4, it turns out that using a fine-grained semantic task as image captioning is the best proxy (highest correlation) of real human judgment.

Fig. 9 shows a captioning example from the COCO images used in the subjective quality evaluation experiment. On the left we show a sample compressed with JPEG with a QF=10, on the center we show the image restored with [14] and on the right we show the original one. It can be observed that the caption of the restored image is capable of describing correctly the image content, on par with the caption obtained on the original image. Instead, the caption of the highly compressed JPEG image is completely unrelated to image content, probably due to object detection errors.



A couple of people sitting next to a christmas tree. | A man riding a wave on a surfboard in the ocean. | A man riding a wave on a surfboard in the ocean.

A teddy bear sitting next to a car. | A dog sitting in the front of a car. | A dog is sitting in a car seat.

A man riding a skateboard on a skate board. | A man riding a skateboard on the street. | A man riding a skateboard on a sidewalk.

Fig. 9. Examples of captions for COCO images used in the subjective quality evaluation. Left column) JPEG compressed with QF=10; Center column) GAN-based restoration from JPEG compressed images with QF=10; right column) original images.

Table 4. Pearson's Correlation coefficient, $\rho(X, Y)$ between No-Reference and captioning based metrics ($x_i \in X$) and MOS ($y_i \in Y$), as defined in Eq. 9 on for a sample set $\mathcal{D}$ sampled from COCO.

Metric	NIQE	BRISQUE	LANBIQUE
$\rho \uparrow$	0.84	0.89	**0.96**

5.2 Results on all distortions

We further show the performance of our approach in full reference image quality assessment on other types of distortion. In this experiment we keep using LIVE dataset, as it contains images corrupted with other processes, such as Gaussian blur, fast-fading, JPEG2000 and white noise, but we add also a recent large scale PieAPP dataset.

5.2.1 Results on LIVE. We repeat the same experiment done for JPEG images on LIVE dataset, firstly considering each distortion separately and then all the distortions together. In Tab. 5 we show the Pearson score for LANBIQUE and several Full-Reference approaches. As we can see, our approach seems to underperform on each distortion except for JPEG, while SSIM and LPIPS are consistent despite the diversity of decaying processes. This is somehow expected, as blur and white noise tend not to harm detection significantly unless they are used with high intensity. Fast fading on the other hand, is to be considered as local distortion. For this reason, objects may not be corrupted at all, thus leading to unchanged detection performances and consequently low correlation scores for our assessment approach. As expected LANBIQUE-NR obtains a lower score than LANBIQUE-NC: in fact LANBIQUE-NC is an upper bound for the No-Reference version since this latter protocol would require a perfect blind restoration method capable of obtaining the reference images to obtain the same score.

Table 5. Pearson's correlation of our approach (Full-Reference and No-Reference) on all distortions present on LIVE compared with other Full-Reference metrics. For the No-Reference approach (LAMBIQUE-NR) fast fading score is not reported since actual State-Of-The-Art restoration approaches perform poorly on this distortion.

	GBLUR	FASTFADING	JP2K	JPEG	WN	TOTAL
PSNR	0.767	0.763	0.83	0.857	0.732	0.752
SSIM	0.886	**0.845**	**0.89**	0.893	**0.951**	0.789
LPIPS	**0.951**	0.836	0.885	0.859	0.910	0.785
LANBIQUE-NC	0.786	0.651	0.787	**0.901**	0.735	**0.792**
LANBIQUE-NR	0.676	-	0.679	0.796	0.667	0.701

However, we experience a totally different scenario when the distortions are evaluated all together. We can see that for each IQA approach we have tested, there is a significant drop in the correlation coefficient with respect to single distortion experiments. We argue this is due to the fact that the scores for single distortion types are well correlated but considering the scores for multiple distortion classes there is a bigger discrepancy between them that leads to a decrease of the total score. On the other hand, our approach does not suffer from this phenomenon, as the performance we measure in these conditions is consistent, if not higher, with single distortions. Moreover, our language based approach slightly overperforms the other measures on the same data and at the same conditions.

5.2.2 Results on PieAPP. Finally, we use a more recent large scale dataset [37]. Prashnani *et al.* collected a very large dataset increasing the number of distortions with respect to existing IQA benchmarks. Moreover, they designed the testing procedure differently. Specifically, instead of collecting multiple subjective scores from a set of users, they rely on the fact that for humans is easier to tell which of two distorted images I_A, I_B is closer to a reference undistorted one I_R. Then images are labelled by the percentage of users that preferred an I_A with respect to I_B. If there is an even split between these two populations, it means that both images are equally different from the reference I_R. Starting from 200 reference images and combining a diverse set of 75 distortions, with a total of 44 distortions in the training set, and 31 in the test set which are distinct from the training set, the PieAPP dataset accounts for a total of 77,280 pairwise comparisons for training (67,200 inter-type and 10,080 intra-type). In Tab. 6 we report results in term of Kendall's Rank Correlation Coefficient: $KRCC = 1/\binom{n}{2} \sum_{i<j} \mathrm{sign}(x_i - x_j)\mathrm{sign}(y_i - y_j)$;

Pearson's Linear Correlation Coefficient (PLCC or $\rho(X, Y)$ as defined in Eq. 9) and Spearman's Rank Correlation (SRCC), $\rho(R(X), R(Y)$ where $R(X)$ are the ranks of sample X.

Interestingly, both image and type of distortions do not overlap between training and testing. In Tab. 6, we show how our LANBIQUE-NC approach (using CIDEr and [11]) ranks with respect to non-learning (top) and learning based (bottom) approaches. We refer to non-learning methods when the algorithm is not relying in any way on any kind of supervision for the IQA task. Our approach exploits learned deep networks and features but those are not the result of training on PieAPP or on any other IQA dataset. Instead, the lower portion of the Table reports methods [7, 8, 24, 29], that are specifically trained to score image similarity. Very interestingly our LANBIQUE-NC approach is consistently better than any non-learned image similarity metric and outperforms all both [7] and [37], with [7] being a close comparison.

Table 6. Evaluation on PieAPP dataset. Column FR indicates if the method is used in a Full-Reference fashion or not. For all metrics higher is better. We report Kendall's Rank Correlation Coefficient (KRCC), Pearson's Linear Correlation Coefficient (PLCC) and Spearman's Correlation Coefficient(SRCC). KRCC is computed for the whole set ($pAB \in [0, 1]$) and for a set for which there is more agreement between human labels ($pAB \notin [.35, .65]$). LANBIQUE-NC has better KRCC with respect to all non-learning based methods and is also better than most of the methods that exploit some sort of supervision to perform IQA.

Method	FR	Learning	KRCC ($pAB \in [0, 1]$)	KRCC ($pAB \notin [.35, 65]$)	PLCC	SRCC
MAE	yes	no	.252	.289	.302	.302
RMSE	yes	no	.289	.339	.324	.351
SSIM	yes	no	.272	.323	.245	.316
MS-SSIM	yes	no	.275	.325	.051	.321
GMSD	yes	no	.250	.291	.242	.297
VSI	yes	no	.337	.395	.344	**.393**
PSNR-HMA	yes	no	.245	.274	.310	.281
FSIMc	yes	no	.322	.377	**.481**	.378
SFF	yes	no	.258	.295	.025	.305
SCQI	yes	no	.303	.364	.267	.360
LANBIQUE-NC	yes	no	**.342**	**.412**	.316	.310
DOG-SSIMc [36]	yes	yes	.263	.320	.417	.464
Lukin et al. [29]	yes	yes	.290	.396	.496	.386
Kim et al. [24]	yes	yes	.211	.240	.172	.252
Bosse et al. [7]	no	yes	.269	.353	.439	.352
Bosse et al. [7]	yes	yes	.414	.503	.568	.537
PieAPP [37]	yes	yes	**.668**	**.815**	**.842**	**.831**

6 CONCLUSION

In this work we propose LANBIQUE, a new approach to evaluate image quality using language models. Existing metrics based on the comparison of the restored image with an undistorted version may give counter-intuitive results. On the other hand, the use of naturalness based scores may in certain cases ranks restored images higher than original ones.

We show that instead of using signal based metrics, semantic computer vision tasks can be used to evaluate results of image enhancement methods. Our claim is that a fine grained semantic computer vision task can be a great proxy for human level image judgement. Indeed we find out that employing algorithms mapping input images to a finer output label space, such as captioning, leads to more discriminative metrics.

LANBIQUE is capable to evaluate the quality of images corrupted by different distortions and its performance is comparable to other image quality assessment methods. Moreover, we have modified our evaluation pipeline to transform our original solution into a No-Reference method and we have demonstrated that it keeps performing fair on standard benchmarks.

Finally, we have tested LANBIQUE an a large scale dataset that contains unknown distortions. Despite the lack of learning and of knowledge on data, our approach outperforms every baseline that does not use learning for the evaluation, and it is comparable to most of the learned approaches on the same data. As a final note, we would like to remark that our approach will continuously improve thanks to the advancement of image captioning and enhancement networks. Indeed, we have shown that without changing the visual features, switching to a better captioning algorithm we get a higher performance. Moreover, being LANBIQUE-NC an upper bound for LANBIQUE-NR, as image enhancers gain quality, the gap between the performance of these two methods will shrink.

ACKNOWLEDGMENTS

This work was partially supported by the European Commission under European Horizon 2020 Programme, grant number 951911 - AI4Media, and grant number 101004545 - ReInHerit.

REFERENCES

[1] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. In *Proc. of ECCV*. Springer, 382–398.

[2] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proc. of CVPR*. 6077–6086.

[3] Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proc. of ACL workshop*. 65–72.

[4] Yochai Blau and Tomer Michaeli. 2019. Rethinking lossy compression: The rate-distortion-perception tradeoff. In *Proc. of ICML*. PMLR, 675–685.

[5] Pietro Bongini, Riccardo Del Chiaro, Andrew D Bagdanov, and Alberto Del Bimbo. 2019. GADA: Generative adversarial data augmentation for image quality assessment. In *Proc. of ICIAP*. Springer, 214–224.

[6] Ali Borji. 2019. Pros and cons of GAN evaluation measures. *Computer Vision and Image Understanding* 179 (2019), 41 – 65. https://doi.org/10.1016/j.cviu.2018.10.009

[7] Sebastian Bosse, Dominique Maniry, Klaus-Robert Müller, Thomas Wiegand, and Wojciech Samek. 2017. Deep neural networks for no-reference and full-reference image quality assessment. *IEEE Transactions on image processing* 27, 1 (2017), 206–219.

[8] H. Chang, M. K. Ng, and T. Zeng. 2014. Reducing Artifacts in JPEG Decompression Via a Learned Dictionary. *IEEE Transactions on Signal Processing* 62, 3 (Feb 2014), 718–728. https://doi.org/10.1109/TSP.2013.2290508

[9] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft COCO captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325* (2015).

[10] Manri Cheon, Sung-Jun Yoon, Byungyeon Kang, and Junwoo Lee. 2021. Perceptual image quality assessment with transformers. In *Proc. of CVPR*. 433–442.

[11] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. 2020. Meshed-memory transformer for image captioning. In *Proc. of CVPR*. 10578–10587.

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *Proc. of CVPR*. 248–255.

[13] Leonardo Galteri, Lorenzo Seidenari, Marco Bertini, and Alberto Del Bimbo. 2017. Deep generative adversarial compression artifact removal. In *Proc. of ICCV*. 4826–4835.

[14] Leonardo Galteri, Lorenzo Seidenari, Marco Bertini, and Alberto Del Bimbo. 2019. Deep universal generative adversarial compression artifact removal. *IEEE Transactions on Multimedia* 21, 8 (2019), 2131–2145.

[15] Leonardo Galteri, Lorenzo Seidenari, Pietro Bongini, Marco Bertini, and Alberto Del Bimbo. 2021. Language Based Image Quality Assessment. In *Proc. of ACM Multimedia Asia*. 1–7.

[16] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Proc. of NIPS*, Vol. 27. 2672–2680.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proc. of CVPR*. 770–778.

[18] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Proc. of NIPS*, Vol. 30. 6629–6640.

[19] Vlad Hosu, Hanhe Lin, Tamas Sziranyi, and Dietmar Saupe. 2020. KonIQ-10k: An ecologically valid database for deep learning of blind image quality assessment. *IEEE Transactions on Image Processing* 29 (2020), 4041–4056.

[20] Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proc. of ICML*. 448–456.

[21] ITU 2012. *Rec. ITU-R BT.500-13 - Methodology for the subjective assessment of the quality of television pictures*. ITU.

[22] Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *Proc. of CVPR*. 4401–4410.

[23] Dong-Wook Kim, Jae Ryun Chung, and Seung-Won Jung. 2019. Grdn: Grouped residual dense network for real image denoising and GAN-based real-world noise modeling. In *Proc. of CVPR Workshops*.

[24] Jongyoo Kim and Sanghoon Lee. 2017. Deep learning of human visual sensitivity in image quality assessment framework. In *Proc. of CVPR*. 1676–1684.

[25] H. Ko, D. Y. Lee, S. Cho, and A. C. Bovik. 2020. Quality Prediction on Deep Generative Images. *IEEE Transactions on Image Processing* 29 (2020), 5964–5979.

[26] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision* 123, 1 (2017), 32–73.

[27] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Proc. of ACL*. 74–81. https://aclanthology.org/W04-1013

[28] Xialei Liu, Joost Van De Weijer, and Andrew D Bagdanov. 2017. Rankiqa: Learning from rankings for no-reference image quality assessment. In *Proc. of ICCV*. 1040–1049.

[29] Vladimir V Lukin, Nikolay N Ponomarenko, Oleg I Ieremeiev, Karen O Egiazarian, and Jaakko Astola. 2015. Combining full-reference image visual quality metrics by neural network. In *Proc. of Human Vision and Electronic Imaging XX*, Vol. 9394. SPIE, 172–183.

[30] Filippo Mameli, Marco Bertini, Leonardo Galteri, and Alberto Del Bimbo. 2021. A NoGAN approach for image and video restoration and compression artifact removal. In *Proc. of ICPR*. 9326–9332.

[31] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. 2012. No-reference image quality assessment in the spatial domain. *IEEE Transactions on image processing* 21, 12 (2012), 4695–4708.

[32] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. 2013. Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters* 20, 3 (2013), 209–212.

[33] Augustus Odena, Vincent Dumoulin, and Chris Olah. 2016. Deconvolution and Checkerboard Artifacts. *Distill* (2016). http://distill.pub/2016/deconv-checkerboard.

[34] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proc. of ACL*. 311–318. https://aclanthology.org/P02-1040

[35] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. 2019. Semantic image synthesis with spatially-adaptive normalization. In *Proc. of CVPR*. 2337–2346.

[36] Soo-Chang Pei and Li-Heng Chen. 2015. Image quality assessment using human visual DOG model fused with random forest. *IEEE Transactions on Image Processing* 24, 11 (2015), 3282–3292.

[37] Ekta Prashnani, Hong Cai, Yasamin Mostofi, and Pradeep Sen. 2018. Pieapp: Perceptual image-error assessment through pairwise preference. In *Proc. of CVPR*. 1808–1817.

[38] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Proc. of NIPS*. 91–99.

[39] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. 2016. Improved techniques for training GANs. In *Proc. of NIPS*, Vol. 29. 2234–2242.

[40] Hamid R Sheikh, Muhammad F Sabir, and Alan C Bovik. 2006. A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Transactions on image processing* 15, 11 (2006), 3440–3451.

[41] Konstantin Shmelkov, Cordelia Schmid, and Karteek Alahari. 2018. How good is my GAN?. In *Proc. of ECCV*. 213–229.

[42] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proc. of CVPR*. 2818–2826.

[43] Hiroki Tomosada, Takahiro Kudo, Takanori Fujisawa, and Masaaki Ikehara. 2021. GAN-based image deblurring using DCT discriminator. In *Proc. of ICPR*. 3675–3681.

[44] Linh Duy Tran, Son Minh Nguyen, and Masayuki Arai. 2020. GAN-based noise model for denoising real images. In *Proc. of ACCV*.

[45] Federico Vaccaro, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. 2021. Fast Video Visual Quality and Resolution Improvement using SR-UNet. In *Proc. of ACM MM*. 1221–1229.
[46] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. CIDEr: Consensus-based image description evaluation. In *Proc. of CVPR*. 4566–4575.
[47] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. 2021. Real-ESRGAN: Training real-world blind super-resolution with pure synthetic data. In *Proc. of ICCV*. 1905–1914.
[48] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. 2018. ESRGAN: Enhanced super-resolution generative adversarial networks. In *Proc. of ECCV workshops*.
[49] Zhou Wang, Alan C Bovik, and Ligang Lu. 2002. Why is image quality assessment so difficult?. In *Proc. of ICASSP*, Vol. 4. IV–3313.
[50] Stefan Winkler. 2009. On the properties of subjective ratings in video quality experiments. In *Proc. of QME*.
[51] Jaeyoung Yoo, Sang-ho Lee, and Nojun Kwak. 2018. Image restoration by estimating frequency distribution of local patches. In *Proc. of CVPR*. 6684–6692.
[52] Junyong You and Jari Korhonen. 2021. Transformer for image quality assessment. In *Proc. of ICIP*. 1389–1393.
[53] Kaihao Zhang, Wenhan Luo, Yiran Zhong, Lin Ma, Bjorn Stenger, Wei Liu, and Hongdong Li. 2020. Deblurring by realistic blurring. In *Proc. of CVPR*. 2737–2746.
[54] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proc. of CVPR*. 586–595.