Language Based Image Quality Assessment

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

ABSTRACT

Evaluation of generative models, in the visual domain, is often performed providing anecdotal results to the reader. In the case of image enhancement, 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 [9, 25, 33]. In this paper we advocate the use of language generation tasks to evaluate the quality of restored images. We show experimentally that image captioning, used as a downstream task, may serve as a method to score image quality. Captioning scores are better aligned with human rankings with respect to signal based metrics 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.

KEYWORDS

image quality enhancement, image captioning, image quality evaluation, GAN, generative models evaluation

ACM Reference Format:
Leonardo Galteri, Lorenzo Seidenari, Pietro Bongini, Marco Bertini, Alberto Del Bimbo. 2021. Language Based Image Quality Assessment. In *ACM Multimedia Asia (MMAsia '21), December 1–3, 2021, Gold Coast, Australia.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3469877.3490605

1 INTRODUCTION & PREVIOUS WORK

In the last years, models able to generate novel images by implicit sampling from the data distribution have been proposed [11]. While these models are extremely appealing, generating for example photo realistic faces [15] or landscapes [23], they are hard to evaluate. 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 [26]. The authors propose the Inception Score (IS), which is obtained applying the Inception model [29] 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 [13]. The authors show that FID is more consistent than Inception Score with increasing disturbances and human judgment. FID performs better as an evaluation metric since it also exploits the statistics of the real images.

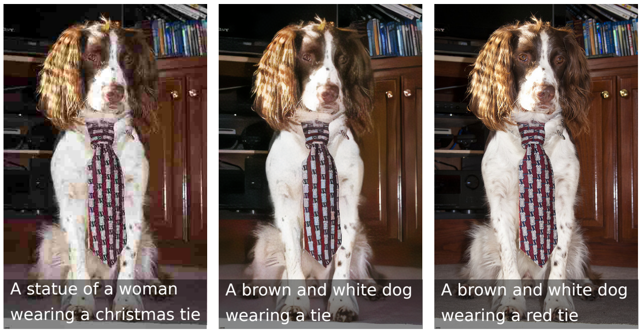

A statue of a woman wearing a christmas tie
A brown and white dog wearing a tie
A brown and white dog wearing a red tie

Figure 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 neck tie". Best viewed in color on computer screen.

Recently [6, 16, 28] have specifically addressed methods to evaluate GANs. In [28] have been proposed two methods that evaluate diversity and quality of generated images using classifiers trained and tested on generated images. In [5] an IQA model is trained with generated images. 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. In [16] it is observed that many existing image quality algorithms do not assess correctly 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.

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).

MMAsia '21, December 1–3, 2021, Gold Coast, Australia
© 2021 Association for Computing Machinery.
ACM ISBN 978-1-4503-8607-4/21/12…$15.00
https://doi.org/10.1145/3469877.3490605

However, these metrics have been often criticized because they are not consistent with human perceived quality of images [31]. 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 tradeoff 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.

Alternatively, no-reference image assessment can be employed. These 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.

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 the obvious drawback of human annotators need and the related cost in terms of time and money in order to rank a high volume of data.

Regarding image enhancement methods, only recently has been proposed to use semantic computer vision tasks to assess image quality. The reason is twofold. On the one hand, images are often processed by algorithms and it is interesting per-se to evaluate the performance of such algorithms on degraded and restored images; to this regard it has to note the MPEG 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 [9, 10, 33].

The main contribution of our work are the following:

- We propose an 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.
- Our evaluation protocol show consistency across different captioning algorithms [2, 7] and language similarity metrics. Interestingly, improving the language generation model also improves the correlation between our score and MOS.
- Experiments shows that our approach 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.

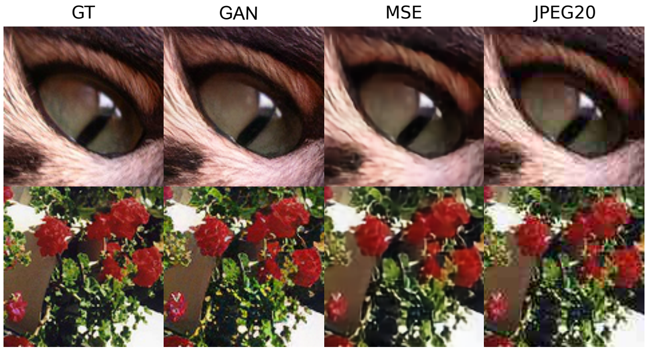


Figure 2: Qualitative comparison of reconstruction methods: GAN produces images more pleasant for the human eye. Best viewed in color on computer screen. GT: original image; JPEG 20: JPEG compression with quality factor 20; MSE: CNN-based restoration using MSE loss and direct training; GAN: GAN-based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

2 IMAGE RESTORATION

Here we formalize the image restoration task. 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 pick a state-of-the art image enhancement method aimed at compression artifact removal, originally presented in [9]. In this work 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 [12]. 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 gridlike patterns possibly stemming from transposed convolutions.

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

$$\mathcal{L}_d = -\log\left(D_\psi\left(I|I^C\right)\right) - \log\left(1 - D_\psi\left(I^R|I^C\right)\right)$$

where I is the uncompressed or high-quality image, I^R is the restored image created by the generator and I^C 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:

$$\mathcal{L}_{adv} = -\log\left(D_\psi\left(I^R|I^C\right)\right) \tag{2}$$

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 and I^R on a feature space by some differentiable function ϕ and taking the Euclidean distance between the two feature representations:

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

In [9] 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.

3 EVALUATION PROTOCOL

Classic full-reference image quality evaluation methods rely on the similarity between an image which has been processed by some enhancement method and a reference undistorted image. GANs are great at filling in high frequency realistic details in image enhancement tasks. Unfortunately this often results in lower performance in full-reference assessment as can be seen in Tab. 4, although the restored images appear as "natural" and pleasant to human evaluators. It is clear from such results that while measuring SSIM and PSNR, optimizing MSE or SSIM losses without adversarial learning is best. For this reason, in [9, 10] 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 at a finer level. 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.

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 which is possibly grammatically and semantically correct, describing the image in detail. We look at performance of a captioning algorithm $\mathcal{A}$ on different versions of a dataset (e.g. COCO): compressed, original and restored. In particular we analyze results from two highly performing captioning methods [2, 7] 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 [24] which is based on a feature extractor pre-trained on ImageNet [8] and then fine-tuned to predict object entities and their attributes using the Visual Genome dataset [17].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 [7] 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 [22], in certain cases including the ordering of words [3], considering n-grams and not just words [18, 30] and the semantic propositional content (SPICE [1]). These metrics evaluate the similarity with respect to a set of reference captions (usually this is five references).

3.1 Subjective evaluation

In this evaluation we assess how images obtained with the selected GAN based restoration method [9] 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 [32]; 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 (a 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 [14].

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

$$e_\mu = \sum_{v \in \mathcal{V}} |\mu^*(v) - \overline{\mu}(v)| \tag{4}$$

and

$$e_{\sigma^2} = \sum_{v \in \mathcal{V}} |\sigma^{2^*}(v) - \overline{\sigma}^2(v)| \tag{5}$$

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

by iterative resampling images until we find e_μ and e_{σ^2} such that $e_\mu \leq 10^{-3}$ and $e_{\sigma^2} \leq 10^{-4}$. Where $\mathcal{V}$ is the set of different version of an image, namely: JPEG compressed at QF=10 (referred to as JPEG 10 in the following), its GAN reconstruction and the original uncompressed image. The selected images contain different subjects, such as persons, 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.

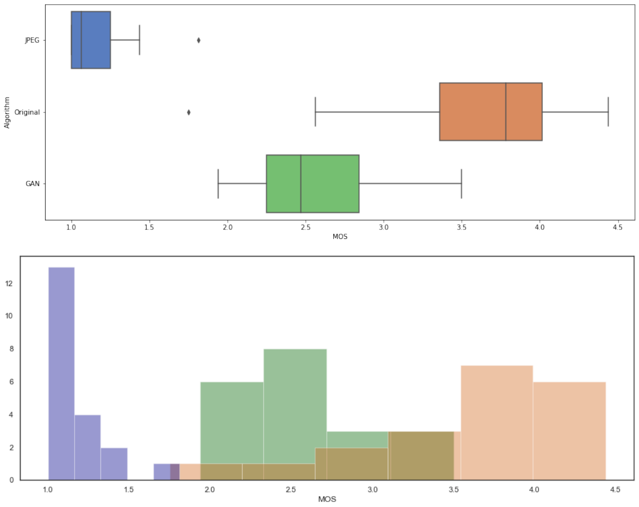


Figure 3: *Top)* Subjective image quality evaluation of original COCO images (orange), heavily compressed JPEG images (blue) and their restored version obtained with the GAN-based approach (green). Restored images are perceived as having a better quality than their compressed versions. *Bottom)* Histograms of MOS scores of the three types of images.

4 RESULTS

In the following, we report results on two datasets: MS-COCO [19] and LIVE [27]. We use COCO, in particular the Karpathy split, since it is the reference benchmark for image captioning, accounting for 5000 images for training and validation each with 5 ground truth sentences per image. LIVE is a widespread benchmark for image quality assessment. LIVE 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.

4.1 Language Based IQA

In Tab. 1 we report results using various captioning metrics. 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 only pleasant to the human eye but recovers details which are also semantically relevant to an 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 a 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 using GAN and captioning as a semantic task (best results highlighted in bold). Captions created from reconstructed images obtain a better score for every metric.

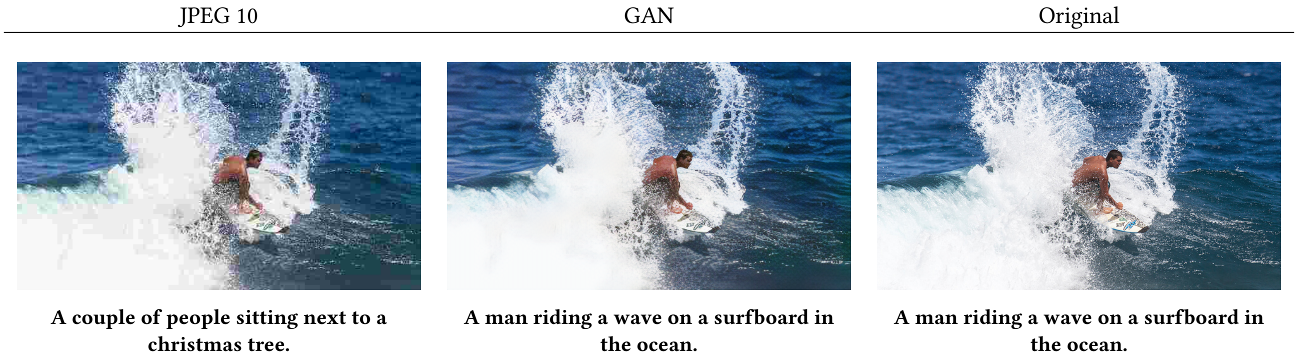
QUALITY	BLEU_1↑	METEOR↑	ROUGE↑	CIDEr↑	SPICE↑
JPEG 10	0.589	0.173	0.427	0.496	0.103
REC 10	**0.730**	**0.253**	**0.527**	**1.032**	**0.189**
JPEG 20	0.709	0.241	0.513	0.937	0.174
REC 20	**0.751**	**0.266**	**0.543**	**1.105**	**0.201**
JPEG 30	0.740	0.258	0.535	1.054	0.194
REC 30	**0.757**	**0.269**	**0.549**	**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. 5 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 [7] outperforms [2] as expected, but interestingly we may observe that the range of CIDEr values of [7] 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. 6 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.

In order to understand better what metric could be used instead of human evaluation we computed the correlation coefficient ρ between BRISQUE [20], NIQE [21], CIDEr and MOS for all versions of the images. As shown in Tab. 2, it turns out that using a fine-grained semantic task as image captioning is the best proxy (highest correlation) of real human judgment.

Fig. 4 show 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 [9] 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.

Figure 4: 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.

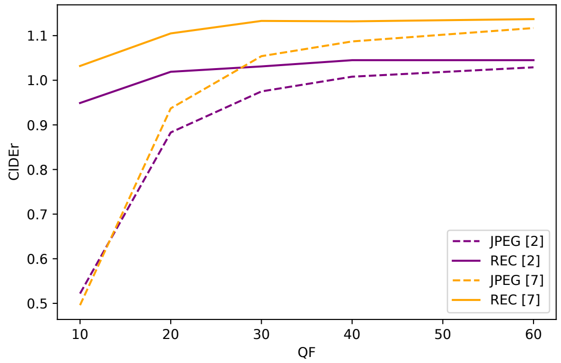


Figure 5: CIDEr scores using [2] and [7] on compressed and restored images for different QFs from MS-COCO.

Table 2: Correlation coefficient between no-reference and captioning based metrics and MOS on COCO.

Metric	ρ
NIQE	0.84
BRISQUE	0.89
CIDEr	**0.96**

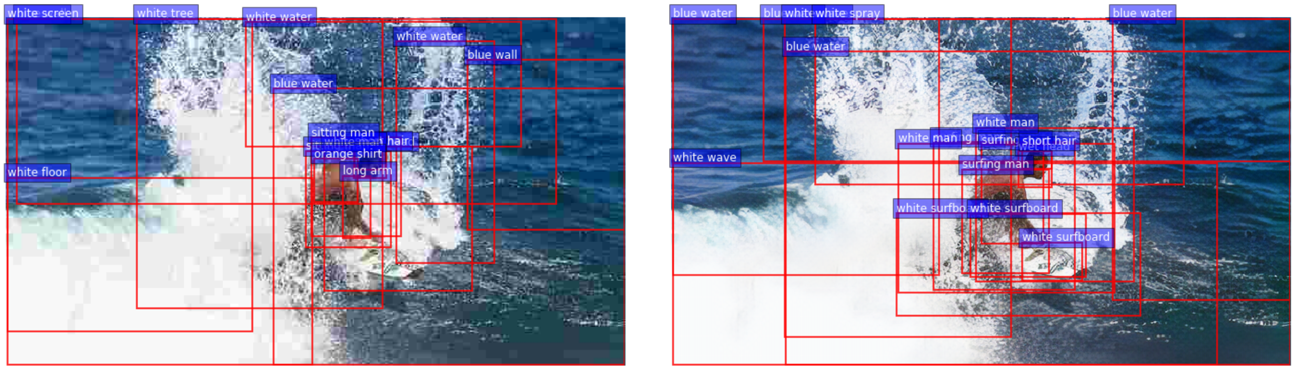
4.2 Comparison with MOS

In Fig. 3 *top)* are reported subjective evaluation results as MOS (Mean Opinion Scores) 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 [9] 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. 3 *bottom)* 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.

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

Metric	**Ours** w/ [7]	**Ours** 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	

We further show that our language based approach correlates with perceived quality using a IQA benchmark test on the LIVE dataset, which contains the opinion scores for each 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. 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 which captioning algorithm is used. Moreover, we observe that the correlation metric significantly improves if we employ a more performing captioner. In this particular 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 to provide detailed and meaningful captions the better we could use the generated sentences to formulate good predictions about the quality of images.



A couple of people sitting next to a Christmas tree.

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

Figure 6: 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. This two indoor details are the one that likely mislead the captioning.

4.3 Comparison with full-reference metrics

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 Tab. 4 we report results on COCO for full-reference indexes. In this setup, we compress the original images at different quality factors and then we restore them with a QF specific artifact removal GAN. We use the uncompressed image generated caption as GT, 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. This is an expected outcome, as in [9] 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. Our approach, using [7], is in line with LPIPS [34]. Unfortunately, LPIPS, as shown in Tab. 3 has low correlation with scores determined by human perceived quality.

4.4 Comparison with no-reference metrics

In certain cases it is not possible to use full reference metrics 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. We report in Tab. 4 the results.

Interestingly, these metrics tend to favor GAN restored images instead of the original uncompressed ones. 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.

Table 4: Evaluation using no-reference and full-reference metrics on MS-COCO. NIQE and BRISQUE rate better GAN images than the ORIGINAL. SSIM always rate restored images worse than compressed. PSNR shows negligible improvement.

QUALITY	NIQE↓	BRISQUE↓	Ours w/ [7] ↑	PSNR ↑	SSIM↑	LPIPS↓
JPEG 10	6.689	52.67	0.542	25.45	0.721	0.305
GAN 10	3.488	17.93	1.118	25.70	0.718	0.144
JPEG 20	5.183	43.99	0.956	27.46	0.796	0.187
GAN 20	3.884	17.85	1.289	27.60	0.784	0.085
JPEG 30	4.474	37.72	1.165	28.61	0.831	0.134
GAN 30	3.601	18.32	1.370	28.81	0.819	0.060
JPEG 40	4.011	33.61	1.260	29.41	0.852	0.105
GAN 40	3.680	18.68	1.424	29.44	0.836	0.048
JPEG 60	3.588	28.15	1.366	30.71	0.880	0.067
GAN 60	3.885	19.45	1.482	30.61	0.862	0.032
ORIGINAL	3.656	21.79	-	-	-	-

5 CONCLUSION

In this work we propose a new idea to evaluate image enhancement methods. 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 rank restored images higher than original ones.

We have shown 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.

We show that employing algorithms mapping input images to a finer output label space, such as captioning, leads to more discriminative metrics. Future work will regard the evaluation of captions provided by humans over compressed and restored images. Moreover, we will take into account the accuracy of captions as a further metric to optimize.

ACKNOWLEDGMENTS

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

REFERENCES

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

[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*.

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

[4] Yochai Blau and Tomer Michaeli. 2019. Rethinking lossy compression: The rate-distortion-perception tradeoff, In Proc. of ICML. *arXiv preprint arXiv:1901.07821*.

[5] Pietro Bongini, Riccardo Del Chiaro, Andrew D Bagdanov, and Alberto Del Bimbo. 2019. GADA: Generative adversarial data augmentation for image quality assessment. In *International Conference on Image Analysis and Processing*. 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] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. 2020. Meshed-memory transformer for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10578–10587.

[8] 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*. Ieee, 248–255.

[9] Leonardo Galteri, Lorenzo Seidenari, Marco Bertini, and Alberto Del Bimbo. 2019. Deep Universal Generative Adversarial Compression Artifact Removal. *Transactions on Multimedia* (2019).

[10] Leonardo Galteri, Lorenzo Seidenari, Marco Bertini, and Alberto Del Bimbo. 2017. Deep Generative Adversarial Compression Artifact Removal. In *Proc. of ICCV*. https://arxiv.org/abs/1704.02518

[11] 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*.

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

[13] 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*.

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

[15] Tero Karras, Samuli Laine, and Timo Aila. 2019. A Style-Based Generator Architecture for Generative Adversarial Networks, In Proc. of CVPR. *CoRR*. arXiv:1812.04948 http://arxiv.org/abs/1812.04948

[16] 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.

[17] 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.

[18] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out* (2004).

[19] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *Proc. of ECCV*.

[20] A. Mittal, A. K. Moorthy, and A. C. Bovik. 2012. No-Reference Image Quality Assessment in the Spatial Domain. *IEEE Transactions on Image Processing* 21, 12 (Dec 2012), 4695–4708. https://doi.org/10.1109/TIP.2012.2214050

[21] 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.

[22] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proc. of ACL*.

[23] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. 2019. Semantic Image Synthesis with Spatially-Adaptive Normalization, In Proc. of CVPR. *CoRR*. arXiv:1903.07291 http://arxiv.org/abs/1903.07291

[24] 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*.

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

[26] Tim Salimans, Ian J. Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. 2016. Improved Techniques for Training GANs. *CoRR* abs/1606.03498 (2016). http://arxiv.org/abs/1606.03498

[27] H. R. Sheikh, Z. Wang, L. Cormack, and A. C. Bovik. 2014. LIVE Image Quality Assessment Database Release 2.

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

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

[30] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proc. of CVPR*.

[31] Z. Wang, A. C. Bovik, and L. Lu. 2002. Why is image quality assessment so difficult?. In *Proc. of ICASSP*. https://doi.org/10.1109/ICASSP.2002.5745362

[32] Stefan Winkler. 2009. On the properties of subjective ratings in video quality experiments. In *Proc. of QME*.

[33] Jaeyoung Yoo, Sang-ho Lee, and Nojun Kwak. 2018. Image Restoration by Estimating Frequency Distribution of Local Patches. In *Proc. of CVPR*.

[34] 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*.