Chapter 3 / Capítulo 3

SAP Superintelligence Series
ISSN: 3144-5624
ISBN (complete work): 978-9915-9851-0-7

Superintelligence Series Volume 5: Intelligent Visual Arts and Audiovisual Engineering: Generative AI, Computational Creativity and Superintelligence
ISBN (this volume): 978-9915-9966-5-3
DOI: 10.62486/978-9915-9966-3-9.Ch03
Pages: 24-35

Check for updates

© 2026 The authors. This is an open access book chapter distributed under the terms of the Creative Commons Attribution (CC BY) 4.0 License.

Classical no-reference image quality metrics and human judgments of artificially generated images

Métricas clásicas de calidad de imagen sin referencia y juicios humanos de imágenes generadas artificialmente

Héctor Aaron Arellano Cruz[1]
Víctor Manuel Martínez García[1]
Yennifer Díaz Romero[1]
Celso Fernando Benitez de La Herrán[1]
Rosa Edilma Garzón González[1]
Alma delia Figueroa Suarez[1]

1 Universidad Autónoma de Sinaloa, Vicerrectoría. Mazatlán, México.

ABSTRACT

Introduction: Artificially generated images are produced in large volumes, but their perceptual quality varies widely and is expensive to measure with human raters. Classical no-reference metrics are a cheap substitute whose validity for generated images is uncertain. Methods: A random sample of 600 images from the AGIQA-3K database was drawn with a fixed seed. The no-reference metrics BRISQUE, NIQE and PIQE were computed for each image and compared with the published mean opinion scores for perceived quality and for text-to-image alignment using Pearson and Spearman correlations with 95% confidence intervals. Results: All three metrics correlated negatively with perceived quality, as expected, at magnitudes between -0.317 and -0.487. NIQE showed the strongest association (r = -0.49) and PIQE the weakest (r = -0.32). Correlations with text-to-image alignment were lower, ranging from -0.239 to -0.370. The mean opinion scores for quality and alignment correlated at r = 0.824. Conclusions: Classical no-reference metrics explain only a moderate share of the variance in human judgments of generated images, with NIQE performing best among the three. They should be treated as coarse quality screens rather than substitutes for human evaluation of generated content.

Keywords: image quality assessment; no-reference metrics; AI-generated images; mean opinion score; BRISQUE; correlation

RESUMEN

Introducción: Las imágenes generadas artificialmente se producen en grandes volúmenes, pero su calidad perceptual varía mucho y es costosa de medir con evaluadores humanos. Las métricas clásicas sin referencia son un sustituto barato cuya validez para imágenes generadas es incierta. Métodos: Se extrajo una muestra aleatoria de 600 imágenes de la base AGIQA-3K con una semilla fija. Se calcularon las métricas sin referencia BRISQUE, NIQE y PIQE para cada imagen y se compararon con las puntuaciones medias de opinión publicadas para calidad percibida y alineación texto-imagen mediante correlaciones de Pearson y Spearman con intervalos de confianza del 95%. Resultados: Las tres métricas correlacionaron negativamente con la calidad percibida, como se esperaba, con magnitudes entre -0,317 y -0,487. NIQE mostró la asociación más fuerte (r = -0,49) y PIQE la más débil (r = -0,32). Las correlaciones con la alineación texto-imagen fueron menores, entre -0,239 y -0,370. Las puntuaciones medias de calidad y alineación correlacionaron en r = 0,824. Conclusiones: Las métricas clásicas sin referencia explican solo una parte moderada de la varianza de los juicios humanos sobre imágenes generadas, con NIQE como la mejor de las tres. Deben tratarse como cribados gruesos de calidad, no como sustitutos de la evaluación humana del contenido generado.

Palabras clave: evaluación de calidad de imagen; métricas sin referencia; imágenes generadas por IA; opinión media; BRISQUE; correlación

INTRODUCTION

Generative models now produce images that are difficult to distinguish from photographs, and the volume of such images has grown far faster than the capacity to judge them. Human subjective assessment remains the reference standard for perceived quality, but it is slow and expensive because it requires many raters and controlled viewing conditions under a protocol such as the ITU recommendation for subjective picture assessment.[1,2] The practical consequence is that most generated images are never evaluated by a human, and the field relies instead on automated metrics that are meant to stand in for opinion.(3) This substitution is consequential: the same automated scores are used to rank generation models, to select training data, and to decide which outputs reach users.

The cheapest and oldest of these substitutes are classical no-reference image quality metrics, which require only the image under test and no pristine reference. They occupy one end of a spectrum: full-reference metrics compare an image against an undistorted original, reduced-reference metrics compare against a compact description of the original, and no-reference metrics judge the image alone.[2,4] For generated images, no-reference assessment is the natural fit, because there is no clean original to compare against; the generated image is the object of interest, and any reference would have to be the prompt, which is text rather than an image. BRISQUE, NIQE and PIQE are the three most widely used no-reference metrics, and all three rest on a shared premise inherited from the natural scene statistics literature: a natural, undistorted image has predictable statistical regularities, and departures from those regularities signal perceptual damage.[5-8] The premise is well supported for photographs and for the familiar distortions of compression, noise, and blur, which is why the metrics remain standard tools more than a decade after they were proposed.[9,10] Their training and evaluation, however, took place on databases of naturally distorted photographs such as LIVE and TID2013, whose content is far removed from synthetic imagery.[11]

The three metrics differ in how they operationalize that shared premise, and the differences matter for how far each can be expected to generalize. BRISQUE learns a mapping from feature statistics to a quality score using a support vector regressor trained on a corpus of images distorted in known ways, which makes it implicitly tuned to the distortion types in that corpus.(5) NIQE avoids training altogether by fitting a multivariate Gaussian to features computed from a corpus of pristine images and scoring new images by their distance from that pristine model, a design its authors call completely blind.(6) PIQE takes a more direct route, segmenting the image into blocks, estimating a local distortion for each block from its contrast statistics, and averaging the result.(7) All three were validated on natural photographs with synthetic distortions, and none was designed with the artifacts of generative models in mind, so the question here is not whether the metrics are sound but whether their assumptions survive the transfer to a new kind of content. (4,9)

Artificially generated images break that premise. Their defects are not the classic distortions but a different class of artifacts, including implausible hands and faces, semantic inconsistencies, and stylistic peculiarities that a natural-scene-statistics model was never designed to recognize. [12,13] Dedicated databases such as AGIQA-3K, AIGCIQA2023 and PKU-AIGIQA-4K were assembled precisely because generated images behave differently under conventional quality models. (12,14,15) Whether the classical no-reference metrics still track human opinion in this new setting is an open empirical question, and the few studies that address it directly are recent and limited in scope.(16)

The question has implications beyond the image processing community. Generated imagery is now embedded in the same pipelines that power advanced artificial intelligence systems, where it serves as training data, as synthetic evaluation data, and as user-facing output, and the ability to filter or rank that imagery automatically determines what those systems learn and what they expose to users.[3,13] If the classical metrics that practitioners reach for by default do not track human judgment, then quality control throughout these pipelines rests on a false assumption, and the consequences propagate to every downstream system that consumes the generated images. Establishing how well those metrics actually predict human opinion is therefore a prerequisite for any credible quality assurance of generative output.

The contribution of this study is an estimate of exactly that validity, obtained on a database built for the purpose and with the metrics computed under a documented, reproducible procedure. Rather than another proposal for a better metric, the study asks whether the tools already in wide use are adequate, and it answers with correlations, confidence intervals, and explained-variance figures rather than with qualitative judgments. The objective of this study was

https://doi.org/10.62486/978-9915-9966-3-9.Ch03

to quantify how strongly BRISQUE, NIQE and PIQE correlate with human mean opinion scores for generated images, using a documented random sample from the AGIQA-3K database.

METHODS

STUDY DESIGN

A cross-sectional correlational study was conducted in which three classical no-reference image quality metrics were computed on a sample of artificially generated images and compared with the mean opinion scores that the publishing database had collected for those same images. The study followed the STROBE reporting framework for observational research. No new human judgments were collected; the design reuses published opinion scores and pairs them with newly computed metric values. This secondary-analysis design trades control over the rating procedure for scale and reproducibility: it would be impractical to re-collect opinion scores at the same breadth, and reusing the published scores keeps the study focused on the metric behavior, which is the object of interest.

DATA SOURCE AND SAMPLE

Images and opinion scores were taken from AGIQA-3K, an open database of 2,982 artificially generated images produced by six generative models, each scored by human raters on perceived quality and on text-to-image alignment.[12] The database was chosen because it is openly licensed, because it separates quality from alignment, and because it spans several generator families rather than a single model, which makes it a reasonable substrate for judging whether a metric generalizes. The database is distributed under the MIT license and includes a comma-separated file with one row per image containing the image identifier, the generation prompt, and the normalized mean opinion scores and their standard deviations for the two constructs. The image files themselves were downloaded from a public mirror of the database.

A random sample of 600 images was drawn with a fixed seed so that the analysis is reproducible; this size yields a margin of error below 0.04 for a Pearson correlation of moderate magnitude, and it keeps the metric computation within the reach of a central processing unit. The sample is large enough that the confidence intervals reported in the results are dominated by the strength of the underlying association rather than by sampling noise, which is what makes the moderate correlations a substantive finding rather than a consequence of imprecision. The six source generators were represented in the sample in proportion to their presence in the full database, so the correlation estimates weight each generator roughly as the full database does.

VARIABLES

The outcome variables were the two mean opinion scores provided by the database: perceived quality and text-to-image alignment, both normalized so that their scale is comparable across generators. The exposure variables were the three no-reference metric scores computed on each image: BRISQUE, NIQE and PIQE. For all three metrics a higher score indicates lower predicted quality, so a negative correlation with opinion scores is the expected direction. The image source generator was recorded as a descriptive covariate. No additional variables were derived from the images beyond the metric scores, because the objective was to evaluate the metrics as they are conventionally computed, without auxiliary features that a practitioner would not have.

MEASUREMENT OF NO-REFERENCE QUALITY

BRISQUE transforms an image into locally normalized luminance coefficients and scores them with a support vector regressor trained on natural and distorted photographs.(5) NIQE fits the same coefficients to a generalized Gaussian model and measures the distance from a model fitted to a corpus of pristine natural images.[6] PIQE divides the image into blocks, classifies each block as noisy, blurry, or clean from local contrast statistics, and aggregates the distorted-block scores.(7) These three span the main design choices in classical blind quality assessment: a

https://doi.org/10.62486/978-9915-9966-3-9.Ch03

trained regressor, a pristine-model distance, and a hand-crafted blockwise heuristic, respectively. (4) All three operate on grayscale luminance, discarding color information, which is a further simplification given that color is part of perceived quality. The three metrics were computed with the pyiqa implementation, which reproduces the original feature and model definitions and runs on a central processing unit. Metric computation was deterministic and was verified by recomputation, with identical results on a repeated run. Because the metrics have no reference image, their scores are absolute rather than relative, and they are typically interpreted as a deviation from naturalness rather than as an error against a specific ground truth.

STATISTICAL ANALYSIS

The association between each metric and each opinion score was estimated with the Pearson product-moment correlation and the Spearman rank correlation, together with the 95% confidence interval for the Pearson coefficient obtained through the Fisher transformation. Pearson captures linear association and is the quantity conventionally reported in image quality studies; Spearman is added because the metrics are not normally distributed and because it is robust to the monotone nonlinearities that arise when a metric saturates at the extremes of the quality range. The coefficient of determination was reported as the squared Pearson correlation to express the share of variance in human opinion explained by each metric. Because the analysis compares three metrics against the same opinion scores, the differences between their correlations were interpreted descriptively rather than tested formally; a formal comparison would require a correction for the strong dependence among the three metrics, which share a common feature basis. All analyses were performed with Python and the SciPy statistics module, and the analysis script was executed twice to confirm that the results are reproducible.

RESULTS

SAMPLE CHARACTERISTICS

The sample comprised 600 images from 6 generative models (Table 1). The mean opinion score for perceived quality was 2.52 with a standard deviation of 0.98, and the alignment score was nearly identical at 2.52 with a standard deviation of 0.97. The three metrics spanned their expected ranges: BRISQUE averaged 35.8 (SD 24.3), NIQE 5.79 (SD 2.99), and PIQE 45.4 (SD 15.5). The dispersion of the metrics was considerable relative to their means, with BRISQUE in particular showing a standard deviation of 24.3 around a mean of 35.8, which reflects the wide range of image quality present in the sample and confirms that the metrics are not saturated at a single value. The two opinion scores were strongly correlated with each other, at r = 0.82, which indicates that perceived quality and prompt alignment overlap but are not the same construct, and justifies treating them as separate outcomes.

Table 1. Descriptive statistics of opinion scores and no-reference metrics in the sample.

Variable	Mean	SD
MOS quality	2.52	0.976
MOS alignment	2.521	0.971
BRISQUE	35.786	24.3
NIQE	5.789	2.989
PIQE	45.369	15.462
Images, n	600	

Note: SD = standard deviation; n = 600 images from six generative models.

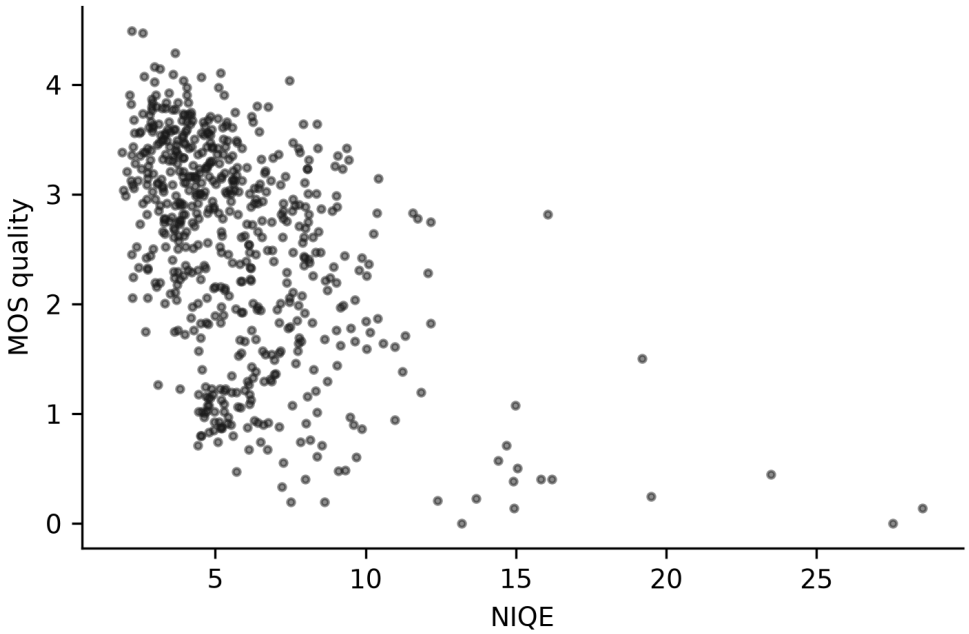
CORRELATION WITH PERCEIVED QUALITY

All three metrics correlated negatively with perceived quality, as expected, and all associations were significant at $p < 0.001$ (Table 2). NIQE showed the strongest association, with a Pearson correlation of −0.49 (95% CI −0.55 to −0.42) and a Spearman correlation of −0.47. BRISQUE

https://doi.org/10.62486/978-9915-9966-3-9.Ch03

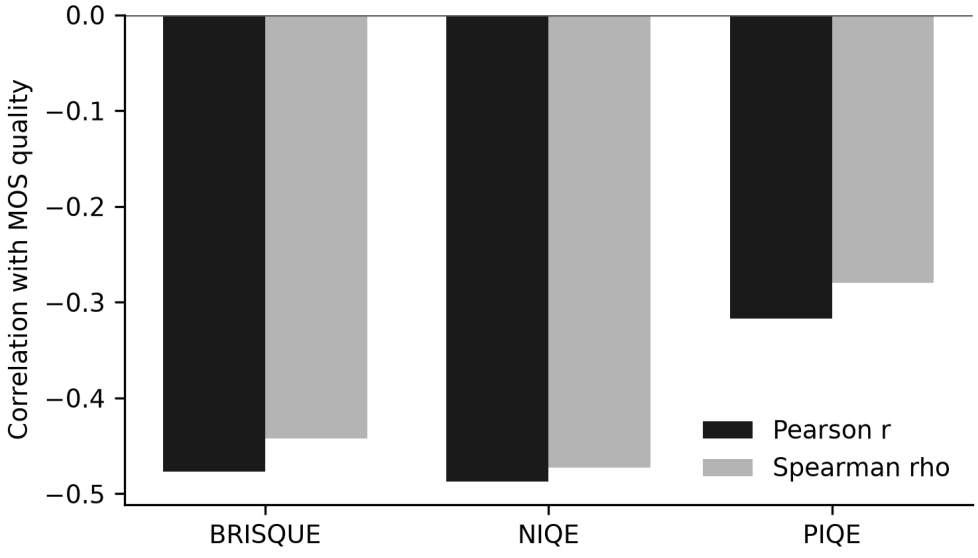
was nearly as strong, at r = −0.48 (95% CI −0.54 to −0.41) and rho = −0.44. PIQE was the weakest of the three, at r = −0.32 (95% CI −0.39 to −0.24) and rho = −0.28. In terms of explained variance, NIQE accounted for roughly one quarter of the variation in human quality judgments, BRISQUE slightly less, and PIQE about one tenth. The confidence intervals of NIQE and BRISQUE overlap, so their difference should be treated as small and possibly not replicable, whereas the interval of PIQE is clearly separated from both, which identifies PIQE as the weakest metric by a meaningful margin. The relationship between the best-performing metric and the quality scores is shown in (Figure 1), and the three metrics are compared side by side in (Figure 2).

Figure 1. Scatter plot of NIQE against mean opinion scores for perceived quality.



Note: Each point is one image; higher NIQE values indicate lower predicted quality.

Figure 2. Correlation of each no-reference metric with mean opinion scores for perceived quality.



Note: Bars show the Pearson r and Spearman rho for BRISQUE, NIQE and PIQE.

The correlations were negative in every case, as the metric definitions require: a higher BRISQUE, NIQE or PIQE score signals a larger departure from naturalness and therefore a lower predicted quality. None of the metrics reversed this sign, and none produced a near-zero association, which confirms that the natural-scene-statistics signal, though attenuated, is not absent from generated images. What is absent is the strong linear relationship that these metrics

https://doi.org/10.62486/978-9915-9966-3-9.Ch03

exhibit on natural distortions; the scatter is wider here, with the best metric leaving the majority of the variance in human opinion unexplained.

Table 2. Correlation of no-reference metrics with mean opinion scores for quality and alignment.

Metric	r (MOS quality)	95% CI	rho (MOS quality)	r (MOS alignment)	rho (MOS alignment)
BRISQUE	-0.477	-0.537 to -0.413	-0.442	-0.348	-0.305
NIQE	-0.487	-0.546 to -0.424	-0.473	-0.37	-0.316
PIQE	-0.317	-0.387 to -0.243	-0.28	-0.239	-0.201

Note: CI = confidence interval; r = Pearson correlation; rho = Spearman correlation; all correlations $p < 0.001$.

CORRELATION WITH TEXT-TO-IMAGE ALIGNMENT

The metrics tracked text-to-image alignment less closely than perceived quality (Table 2). NIQE again led, with $r = -0.37$ and rho = −0.32, followed by BRISQUE at $r = -0.35$ and PIQE at $r = -0.24$. The rank order of the three metrics was the same for alignment as for quality, but every correlation was lower in absolute value, which indicates that the metrics capture perceptual fidelity better than semantic correspondence between prompt and image. The reduction was most pronounced for the metric with the weakest overall signal, PIQE, whose alignment correlation fell to −0.24, and smallest for NIQE, which retained an alignment correlation of −0.37. The strong correlation between the two opinion constructs ($r = 0.82$) means that part of the alignment association is carried through shared variance with quality, but the consistently lower alignment coefficients show that the metrics do not reach the semantic content of the images.

Taken together, the results describe a consistent picture rather than a set of isolated coefficients. Every metric was negative with every outcome, the ranking of the metrics was identical across outcomes, and the quality coefficients exceeded the alignment coefficients in every case. This internal consistency argues that the estimates reflect a stable property of the metrics rather than sampling variation, and it is the stability of the ranking, more than the exact magnitudes, that is most likely to replicate in other datasets.

DISCUSSION

MAIN FINDINGS

All three classical no-reference metrics predicted human judgments of generated images in the expected direction, but only to a moderate degree. NIQE was the strongest, explaining roughly one quarter of the variance in perceived-quality scores, BRISQUE was a close second, and PIQE was markedly weaker, explaining about one tenth. The same ranking held for text-to-image alignment, where every correlation was lower still. No metric approached the strength of agreement that would justify replacing human raters; the best of the three is at most a coarse screen for perceived quality. The negative sign of every correlation is itself a meaningful result, because it confirms that the metrics are not simply random on this content and that their naturalness signal carries a genuine, if partial, correspondence to what humans see. The finding is at once reassuring and cautionary: the classical metrics are not useless on generated images, but they are not sufficient either.

INTERPRETATION AND COMPARISON WITH OTHER STUDIES

The moderate correlations match what the few recent studies that have tested classical metrics on generated content reported. Sridhar and colleagues found that non-reference metrics perform inconsistently on generated images and argued for task-specific scoring, a conclusion this study supports with a larger and better-controlled sample.(16) Chen and colleagues likewise found that subjective studies and objective metrics diverge for generated images and that conventional metrics underrepresent the defects raters actually notice.(13) The present estimate of a Pearson correlation around 0.49 for NIQE places the classical metrics below the learned models that have since been developed for generated images, which typically reach correlations above 0.8 on the same databases.(17-19)

https://doi.org/10.62486/978-9915-9966-3-9.Ch03

The comparison with the classical metrics' own historical performance is even starker. On databases of naturally distorted photographs, BRISQUE and its contemporaries were typically reported with Spearman rank correlations above 0.9, and even the simpler DCT-based predecessors reached the 0.8 range.[11,20] The drop from those levels to the values observed here, on the order of 0.28 to 0.47, quantifies how much of the metrics' validity was tied to the natural-distortion setting in which they were designed. This is not a criticism of the metrics, which were never claimed to cover synthetic content, but it is a reminder that a metric validated in one domain cannot be assumed to transfer to another, and that the transfer gap can be large enough to change a decision from reliable to unreliable.(2,4)

The gap between classical and learned metrics has a clear mechanistic explanation. BRISQUE, NIQE and PIQE encode a statistical model of natural photographs, and their features respond to the luminance and contrast disruptions that characterize compression, noise, and blur.[10,20] Generated images carry a different class of artifacts, including anatomically implausible details and semantic inconsistencies that do not perturb natural scene statistics in the same way.[12,14] A metric that measures distance from a pristine natural photograph therefore registers only part of what makes a generated image look wrong. The lower correlations with alignment follow the same logic: alignment depends on semantic content that a purely statistical metric cannot see at all.(13,15)

This mechanistic account also explains the broader trajectory of the field. The hand-crafted natural-scene-statistics features that define the classical era gave way to learned features extracted by deep networks precisely because the learned representations could capture semantic and perceptual structure that the statistical features could not.[21,22] The classical metrics remain useful because they need no training data and no reference, but their representational ceiling was reached early, and the results here show where that ceiling lies for synthetic content. In this sense the study is less an evaluation of three particular algorithms than a measurement of the limits of an entire, and once dominant, approach to quality assessment. Text-to-image alignment is evaluated by comparing the image against a text prompt, which requires a joint understanding of language and vision that is fundamentally outside the scope of a luminance-statistics metric. The fact that the alignment correlations are nevertheless nonzero is explained by the overlap between the two opinion constructs rather than by any semantic capability of the metrics: an image that is severely degraded perceptually is also, in practice, a poor match to its prompt, so part of the alignment signal rides along with quality. Once that shared component is accounted for, the residual alignment signal, which is the semantic part, is essentially invisible to all three metrics.

The finding that NIQE outperforms BRISQUE and PIQE is also interpretable. NIQE is the most purely statistical of the three, fitting a general model of pristine natural images without a distortion-specific training stage, whereas BRISQUE is trained on a specific distortion corpus and PIQE relies on a local blockwise heuristic.(5-7) On generated images, the distortion-specific assumptions of BRISQUE and the local heuristic of PIQE appear to generalize less well than NIQE's broader naturalness measure, which is consistent with the opinion-unaware philosophy that NIQE introduced and that later deep models adopted.[23,24] The result is also a useful corrective to the default of using BRISQUE or PIQE in applied settings: in the domain of generated images, the metric chosen matters, and the common assumption that all no-reference metrics behave alike does not hold here.

There is also a ceiling on what any of these metrics can achieve, independent of their individual designs. All three read a single image and score it against a statistical model of naturalness, which means they can only see the properties that naturalness captures. Human raters, by contrast, judge generated images on semantic plausibility, on whether the content matches the prompt, and on aesthetic qualities that have little to do with natural scene statistics.[12,25] A metric that cannot see these properties cannot predict them, which places an upper bound on the attainable correlation that is well below what learned models, which ingest semantic features, can reach.[17,18] The moderate correlations reported here should therefore be read not as a failure of the classical metrics but as a statement of the ceiling imposed by their representational assumptions.

GENERALIZABILITY

Within the limits of a single database, the results are likely to generalize to the broader problem of evaluating generated images with classical tools, because the underlying mismatch between natural-scene-statistics models and synthetic artifacts is a property of the metrics,

https://doi.org/10.62486/978-9915-9966-3-9.Ch03

not of the particular generators sampled.[8,12] The same mismatch has been observed outside computer vision, where classical no-reference metrics were applied to medical images produced by generative adversarial networks and found to need human validation.(28) The practical implication for practitioners is straightforward: BRISQUE, NIQE and PIQE can rank a batch of generated images roughly by quality, and NIQE should be preferred when only a classical metric is available, but none should be reported as a measure of human-perceived quality or of prompt alignment.

For research and industrial pipelines that require reliable quality assessment, learned metrics trained on generated-image databases remain the appropriate tool, and even aesthetic preference is better captured by learned scores tuned to human preference than by statistical naturalness.(21,22,25,29,30) The classical metrics retain value as reference-free, training-free baselines against which those learned models can be benchmarked, a role they have played since the early deep quality models appeared.[22,30] A sensible division of labor follows from the results: classical metrics can provide a first, inexpensive pass that flags the worst generations and orders a batch for human review, while learned metrics and human raters handle the decisions that require semantic or aesthetic judgment. The findings also carry a caution for the many applied studies that report BRISQUE or PIQE as a quality indicator for generated content: such numbers describe distance from naturalness, not quality as a human would perceive it, and conflating the two will mislead any downstream decision that depends on quality.

A concrete recommendation follows from the ranking. When a classical no-reference metric is needed and only one can be run, NIQE is the justified choice for generated images on the evidence here, because it outperformed the alternatives on both outcomes and does not rely on a distortion corpus that may not match the content. When computation or interpretability favors a blockwise score, PIQE remains the weakest option and should be treated accordingly. These recommendations are conditional on the database and generators studied, but the consistency of the ranking across the two outcome constructs gives more confidence in it than a single comparison would.

IMPLICATIONS FOR QUALITY ASSURANCE OF GENERATIVE SYSTEMS

The results bear directly on how generated content should be quality-controlled in production systems. A pipeline that uses BRISQUE or PIQE as a filter passes and rejects images on a signal that explains only a small fraction of human quality judgment, so many poor images will pass and many good ones will be rejected. The stronger but still moderate performance of NIQE does not change this conclusion; it only makes NIQE the least bad choice among classical options. For systems whose output is user-facing, the findings argue for reserving human or learned-model judgment for the final acceptance decision and using classical metrics only as an inexpensive pre-filter that flags gross defects.

The implications extend to the governance of generated content. If automated quality metrics are used to decide whether generated images meet a quality bar, whether they are suitable for distribution, or whether they should be labeled as synthetic, then the validity of those metrics is a policy question as much as a technical one. A metric that systematically misses semantic defects would give a false assurance of quality, and a metric tuned to naturalness would penalize stylized but perfectly adequate generations. The findings here suggest that any governance framework relying on classical no-reference scores would be relying on a proxy that human judgment does not fully endorse, and that such frameworks should either adopt learned, semantically aware metrics or keep a human in the loop.[3,29] The classical metrics, in short, are useful instruments whose scope is narrower than the uses to which they are often put, and recognizing that scope is the first step toward deploying them responsibly.

CONCLUSIONS

The classical no-reference metrics BRISQUE, NIQE and PIQE correlated negatively and significantly with human mean opinion scores of generated images, confirming that they retain some validity beyond natural photographs. NIQE was the strongest predictor of perceived quality, followed closely by BRISQUE, while PIQE was substantially weaker, and all three tracked text-to-image alignment less well than perceptual quality. None of the three explained more than a quarter of the variance in human judgments, so they should be used as coarse, training-free

https://doi.org/10.62486/978-9915-9966-3-9.Ch03

quality screens rather than as substitutes for human evaluation of generated content. The findings establish both the continued usefulness and the clear ceiling of the natural-scene-statistics approach, and they point to learned, semantically aware metrics as the appropriate tool wherever quality assessment must track human perception closely.

https://doi.org/10.62486/978-9915-9966-3-9.Ch03

REFERENCES

International Telecommunication Union. Methodology for the subjective assessment of the quality of television pictures (Recommendation ITU-R BT.500-14) [Internet]. 2019 [cited 2026 Oct 25]. Available from: https://www.itu.int/rec/R-REC-BT.500
Zhai G, Min X. Perceptual image quality assessment: a survey. Sci China Inf Sci. 2020;63(11):211301. doi: 10.1007/s11432-019-2757-1
Zhang Z, Zhou Y, Li C, Zhao B, Liu X, Zhai G. Quality Assessment in the Era of Large Models: A Survey. ACM Trans Multimedia Comput Commun Appl. 2025;21(7):1-31. doi: 10.1145/3722559
Shahid M, Rossholm A, Lövström B, Zepernick HJ. No-reference image and video quality assessment: a classification and review of recent approaches. J Image Video Proc. 2014;2014(1):40. doi: 10.1186/1687-5281-2014-40
Mittal A, Moorthy AK, Bovik AC. No-Reference Image Quality Assessment in the Spatial Domain. IEEE Trans on Image Process. 2012;21(12):4695-4708. doi: 10.1109/tip.2012.2214050
Mittal A, Soundararajan R, Bovik AC. Making a "Completely Blind" Image Quality Analyzer. IEEE Signal Process Lett. 2013;20(3):209-212. doi: 10.1109/lsp.2012.2227726
Venkatanath N, Praneeth D, Maruthi Chandrasekhar Bh, Channappayya SS, Medasani SS. Blind image quality evaluation using perception based features. 2015 Twenty First National Conference on Communications (NCC). 2015:1-6. doi: 10.1109/ncc.2015.7084843
Ruderman DL. The statistics of natural images. Network: Computation in Neural Systems. 1994;5(4):517-548. doi: 10.1088/0954-898x_5_4_006
Saad MA, Bovik AC, Charrier C. Blind Image Quality Assessment: A Natural Scene Statistics Approach in the DCT Domain. IEEE Trans on Image Process. 2012;21(8):3339-3352. doi: 10.1109/tip.2012.2191563
Moorthy AK, Bovik AC. Blind Image Quality Assessment: From Natural Scene Statistics to Perceptual Quality. IEEE Trans on Image Process. 2011;20(12):3350-3364. doi: 10.1109/tip.2011.2147325
Ponomarenko N, Jin L, Ieremeiev O, Lukin V, Egiazarian K, Astola J, et al. Image database TID2013: Peculiarities, results and perspectives. Signal Processing: Image Communication. 2015;30:57-77. doi: 10.1016/j.image.2014.10.009
Li C, Zhang Z, Wu H, Sun W, Min X, Liu X, et al. AGIQA-3K: An Open Database for AI-Generated Image Quality Assessment. IEEE Trans Circuits Syst Video Technol. 2024;34(8):6833-6846. doi: 10.1109/tcsvt.2023.3319020
Chen J, Shao F, Chen H, Wang X, Guo H, Jiang Q. Quality Evaluation of AI-Generated Images: Subjective Study and Objective Methodology. IEEE Trans Multimedia. 2026;28:57-70. doi: 10.1109/tmm.2025.3618562
Wang J. AIGCIQA2023: A Large-scale Image Quality Assessment Database for AI Generated Images: from the Perspectives of Quality, Authenticity and Correspondence [Internet]. 2023 [cited 2026 Oct 25]. Available from: https://arxiv.org/abs/2307.00211
Yuan J. PKU-AIGIQA-4K: A Perceptual Quality Assessment Database for Both Text-to-Image and Image-to-Image AI-Generated Images [Internet]. 2024 [cited 2026 Oct 25]. Available from: https://arxiv.org/abs/2404.18409
Sridhar Y, Gajendra JM, Srinivasalu P, Srinivasappa M. Evaluating Non-Reference Image Quality Metrics for AI-Generated Images: A Novel Approach. EJAI. 2025;4(5):18-25. doi: 10.24018/ejai.2025.4.5.70
Zhou T, Tan S, Zhou W, Luo Y, Wang YG, Yue G. Adaptive Mixed-Scale Feature Fusion Network for Blind AI-Generated Image Quality Assessment. IEEE Trans on Broadcast. 2024;70(3):833-843. doi: 10.1109/tbc.2024.3391060
Fu J, Zhou W, Jiang Q, Liu H, Zhai G. Vision-Language Consistency Guided Multi-Modal Prompt Learning for Blind AI Generated Image Quality Assessment. IEEE Signal Process Lett. 2024;31:1820-1824. doi: 10.1109/lsp.2024.3420083
Wang J, Duan H, Zhai G, Min X. Quality Assessment for AI Generated Images With Instruction Tuning. IEEE Trans Multimedia. 2026;28:2404-2419. doi: 10.1109/tmm.2026.3651009
Saad MA, Bovik AC, Charrier C. A DCT Statistics-Based Blind Image Quality Index. IEEE Signal Process Lett. 2010;17(6):583-586. doi: 10.1109/lsp.2010.2045550
Yang X, Li F, Liu H. A Survey of DNN Methods for Blind Image Quality Assessment. IEEE Access. 2019;7:123788-123806. doi: 10.1109/access.2019.2938900
Bosse S, Maniry D, Muller KR, Wiegand T, Samek W. Deep Neural Networks for No-Reference and Full-Reference Image Quality Assessment. IEEE Trans on Image Process. 2018;27(1):206-219. doi: 10.1109/tip.2017.2760518
Ni Z, Liu Y, Ding K, Yang W, Wang H, Wang S. Opinion-Unaware Blind Image Quality Assessment Using Multi-Scale Deep Feature Statistics. IEEE Trans Multimedia. 2024;26:10211-10224. doi: 10.1109/tmm.2024.3405729
Sinno Z, Caramanis C, Bovik AC. Towards a Closed Form Second-Order Natural Scene Statistics Model. IEEE Trans on Image Process. 2018;27(7):3194-3209. doi: 10.1109/tip.2018.2817740
Li D. Playground v2.5: Three Insights towards Enhancing Aesthetic Quality in Text-to-Image Generation [Internet]. 2024 [cited 2026 Oct 25]. Available from: https://arxiv.org/abs/2402.17245
Pin SHD, Amirshahi SA. Individual Differences in Subjective Image Quality Assessment: A Bayesian Mixed-Effects Modeling Approach. IEEE Access. 2025;13:127504-127517. doi: 10.1109/access.2025.3579016
Min X, Gao Y, Cao Y, Zhai G, Zhang W, Sun H, et al. Exploring Rich Subjective Quality Information for Image Quality Assessment in the Wild. IEEE Trans Circuits Syst Video Technol. 2025;35(8):7778-7791. doi: 10.1109/tcsvt.2025.3544659
Treder MS, Codrai R, Tsvetanov KA. Quality assessment of anatomical MRI images from generative adversarial networks: Human assessment and image quality metrics. Journal of Neuroscience Methods. 2022;374:109579.

https://doi.org/10.62486/978-9915-9966-3-9.Ch03

doi: 10.1016/j.jneumeth.2022.109579
An B, Zhou H, Peng P, Zhang L, Ren S, Li J, et al. A survey of learning-based no-reference image quality assessment. The International Conference Optoelectronic Information and Optical Engineering (OIOE2024). 2025;13513:124. doi: 10.1117/12.3045771
Hosu V, Lin H, Sziranyi T, Saupe D. KonIQ-10k: An Ecologically Valid Database for Deep Learning of Blind Image Quality Assessment. IEEE Trans on Image Process. 2020;29:4041-4056. doi: 10.1109/tip.2020.2967829

https://doi.org/10.62486/978-9915-9966-3-9.Ch03