Mahesh B et al., Vol 2(2), pp 63-87
https://doi.org/10.55306/CJDTES.2025.020201
e-ISSN: 3048-8389
Ci-STEM Journal of Digital Technologies and Expert Systems

A Comprehensive Framework for Perceptual Quality Assessment in AI-Based Image Compression

Mahesh. Boddu[1], Soumitra Kumar Mandal[2]
[1]Department of Electrical Engineering, Maulana Abul Kalam Azad University of Technology, Kolkata, West Bengal, India
[2]Department of Electrical Engineering, National Institute of Technical Teachers' Training & Research (NITTTR), Kolkata, West Bengal, India.
[1]email: bmaheshecse@gmail.com, [2]email: skmandal@nitttrkol.ac.in.

Open Access
Research Article
Received: 30/05/2025
Accepted: 16/07/2025
Published: 23/08/2025

Corresponding author email:
bmaheshecse@gmail.com

Citation:
Mahesh B et al., "A Comprehensive Framework for Perceptual Quality Assessment in AI-Based Image Compression" Ci-STEM Journal of Digital Technologies and Expert Systems, Vol. 2(2), pp.63-87, 2025, doi: 10.55306/CJDTES.2025.020201

Copyright:
©2025 Mahesh B et al.,
This is an open-access article distributed under the terms of the Creative Commons Attribution License which grants the right to use, distribute, and reproduce the material in any medium, provided that proper attribution is given to the original author and source, in accordance with the terms outlined by the license.
(https://creativecommons.org/licenses/by/4.0/)
.

Published By:
Ci-STEM Global Services Foundation, India.

Abstract:
Perceptual quality assessment has become the central bottleneck in the design, evaluation, and deployment of artificial-intelligence-based image compression systems. Classical signal-fidelity measures such as mean squared error (MSE) and peak signal-to-noise ratio (PSNR) correlate poorly with human judgment. Learned codecs, meanwhile, increasingly optimize directly for structural, perceptual, and generative objectives. How those objectives behave under the rate–distortion–perception trade-off is still not well understood. This paper presents a critical review of the field spanning 1970 to 2023, drawing on 75 peer-reviewed and archival sources selected through a PRISMA-informed screening protocol. We trace how image compression and quality assessment evolved together, from Shannon's rate-distortion theory through DCT- and wavelet-based standards to learned nonlinear transform coding and generative codecs. The mathematical foundations of rate–distortion theory, structural similarity, information fidelity, and the rate–distortion–perception function are then formalized, and a taxonomy is presented covering classical, structural, perceptual, no-reference, and AI-based quality metrics. Learned and generative compression architectures are examined in detail, hyperprior, autoregressive, transformer, and diffusion-based designs among them. Building on this synthesis, we propose a unified evaluation framework that folds pixel, structural, perceptual, semantic, and task-fidelity dimensions into a single adaptively weighted assessment protocol suited to modern AI-based codecs. The paper also catalogues the subjective datasets and protocols that underlie the field and presents comparative tables of representative metrics and codecs. It closes by identifying six research gaps, nine future research questions, and a short-, medium-, and long-term roadmap. The review is meant to serve as both a technical reference and an agenda-setting document for researchers, standards bodies, and practitioners working where learned compression meets perceptual quality science.

Keywords:
image compression, perceptual quality assessment, rate–distortion–perception trade-off, learned image compression, generative compression, structural similarity, deep perceptual metrics, systematic literature review, subjective image quality, no-reference IQA.

1. INTRODUCTION AND HISTORICAL BACKGROUND

1.1 Introduction

Few technologies in modern computing are as heavily deployed as image compression. Every photograph stored on a smartphone, every frame streamed by a video platform, and every image sent across a bandwidth-constrained network passes through some form of lossy or lossless encoding. For more than three decades the dominant paradigm was set by the discrete cosine transform (DCT)-based JPEG standard [1] and its wavelet-based successor JPEG2000 [2], [3]. Both apply a fixed, hand-engineered linear transform, then quantization, then entropy coding. Performance was measured almost exclusively in terms of mean squared error (MSE) or its logarithmic counterpart, peak signal-to-noise ratio (PSNR). The codecs that resulted are computationally efficient, standardized, and broadly interoperable. Their distortion criterion, however, tracks only weakly with how a human observer actually perceives reconstruction quality [4].

Two developments over the last decade have reinforced each other. The first is learned image compression, in which an encoder–decoder pair and an entropy model are trained jointly and end-to-end on large image corpora. A proof of concept [5] has since grown into a whole family of architectures. Hyperprior models [6], joint autoregressive and hierarchical priors [7], channel-wise context models [8], attention-based Gaussian mixture models [9], checkerboard context models [10], and unevenly grouped space-channel models such as ELIC [11] all now match or exceed the rate–distortion performance of the best conventional codecs, VVC/H.266 [12] included. A second development follows from the first: the criteria used to train and evaluate these codecs have expanded well beyond MSE. Structural similarity (SSIM) and its multi-scale variant [13], [14], information-fidelity criteria [15], deep-feature perceptual metrics such as LPIPS [16] and DISTS [17], and fully generative objectives built on adversarial [18], [19] or diffusion-based [20], [21] decoders have all been proposed as training losses, as evaluation metrics, or as both. “Quality” has consequently become a plural, contested, and architecture-dependent notion.

This pluralism is more than a benchmarking inconvenience. It reflects a real theoretical tension, formalized by Blau and Michaeli [22], [23]. They showed that classical distortion (MSE, for instance) and “perceptual quality” are generally at odds at any fixed rate, where perceptual quality means statistical closeness of the reconstructed image distribution to the distribution of natural images. Forcing perceptual quality to be high provably raises the achievable distortion; there is no free lunch. Generative codecs such as HiFiC [19] and GAN-based extreme compression [18] trade numerical fidelity for perceptual realism by design. A human observer may well prefer their reconstructions even though, pixel for pixel, those reconstructions diverge from the original more than an MSE-optimized one would. For such codecs, an evaluation framework that reports only PSNR or SSIM is worse than incomplete. It can actively mislead the reader about what the codec was built to do.

Against this backdrop, the present paper undertakes a systematic and critical review of perceptual quality assessment in the specific context of AI-based image compression. Several prior surveys have covered learned image compression architectures in isolation, and others have covered image quality assessment (IQA) metrics in isolation. Comparatively few have synthesized the two literatures together, mapped rate–distortion–perception theory onto practical evaluation protocols, or proposed an integrated, multi-dimensional evaluation framework suited to the current generation of generative and semantic codecs. This review is designed to fill that gap. Its contributions are fivefold:

1) A historically grounded synthesis tracing how compression technology and quality assessment methodology co-evolved, from Shannon's 1948–1959 information-theoretic foundations through the standards era to the learned and generative era (Section 2.1).

2) A rigorous mathematical treatment of rate–distortion theory, structural and information-theoretic fidelity criteria, and the rate–distortion–perception function, with every equation traceable to its original published source (Section 2.2).

3) An explicit and reproducible systematic literature review protocol covering research questions, database and search strategy, inclusion and exclusion criteria, a PRISMA-informed screening pipeline, a quality-assessment checklist, and a data-extraction procedure (Section 2.3).

4) A structured taxonomy and critical comparison of classical, structural, information-theoretic, no-reference, and AI-based quality metrics, and of learned and generative compression architectures together with their training losses and known failure modes (Sections 3.1-3.2).

5) A unified, multi-dimensional evaluation framework integrating pixel, structural, perceptual, semantic, and task fidelity, accompanied by a catalogue of subjective datasets (Section 4.1), comparative tables (Section 4.2), identified research gaps Section 5.1), future research questions (Section 5.2), and a phased research roadmap (Section 5.3).

The paper is organized as follows. Section 1.2 reviews the historical evolution of image compression and quality assessment. Section 2.1 develops the mathematical and computational foundations, and Section 2.2 details the systematic review methodology. Section 2.3 presents the taxonomy of quality metrics, and Section 3.1 reviews learned and generative compression. Section 3.2 introduces the proposed unified framework. Section 4.1 catalogues datasets and subjective evaluation protocols, Section 4.2 presents the comparative analysis, and Section 5.1 discusses research gaps. Section 5.2

formulates future research questions, Section 5.3 presents the research roadmap, and Section 5.4 concludes.

1.2 Historical Evolution of Image Compression and Quality Assessment

A. The Information-Theoretic Era (1948–1991)

Every lossy image compression scheme rests, ultimately, on Shannon's foundational work. His 1948 paper establishing information theory [24] introduced entropy as a fundamental limit on lossless representation. His 1959 paper on coding theorems for a discrete source with a fidelity criterion [25] then introduced rate–distortion theory, the formal characterization of the minimum rate R(D) needed to represent a source within an average distortion D under a specified single-letter distortion measure. Berger's 1971 monograph [26] and the standard treatment in Cover and Thomas [27] extended and systematized it for continuous sources, Gaussian channels, and multiterminal settings. It remains the mathematical bedrock against which every subsequent compression scheme, hand-engineered or learned, is measured. Goyal's survey of transform coding [28] later explained why block transforms such as the DCT approximate the theoretically optimal Karhunen–Loève transform for stationary Gaussian sources, and in doing so supplied the theoretical justification for the transform-quantize-entropy-code pipeline that would dominate compression practice for the next four decades.

B. The Standards Era (1992–2012)

Wallace's 1992 formalization of the JPEG standard [1] put transform coding to work at industrial scale: 8×8 block DCT, uniform scalar quantization guided by perceptually motivated quantization tables, zig-zag scanning, and Huffman entropy coding. Its practical success was enormous. Even so, the block-based structure produces characteristic blocking and ringing artifacts at low bit rates. The fidelity criterion behind its quantization tables amounted at best to a coarse approximation of human contrast sensitivity, never a learned or rigorously validated perceptual model. JPEG2000 [2], [3] addressed many of these structural limitations by replacing the block DCT with a discrete wavelet transform, embedded bit-plane coding (EBCOT), and rate-distortion optimized truncation. Degradation became smoother and compression efficiency improved, particularly at low bit rates. Computational cost rose sharply in exchange, and industry adoption stayed confined to specialized domains such as digital cinema and medical imaging. More recent conventional designs such as JPEG XL [29] revisit the same transform-coding foundations with modernized entropy coding and broader colour and dynamic-range support, but they remain squarely within the hand-engineered paradigm.

The video coding standards lineage developed in parallel. H.264/AVC [30], H.265/HEVC [31], and eventually H.266/VVC [12] refined block-based hybrid transform coding with increasingly sophisticated intra- and inter-prediction, larger and more flexible partitioning, and improved entropy coding (CABAC). Each generation targeted roughly a 50% bit-rate reduction over its predecessor at equal subjective quality. The meaning of “equal subjective quality” has historically been settled by formal subjective testing protocols such as ITU-R BT.500 [32] and ITU-T P.910 [33]. Both documents specify controlled viewing conditions, double-stimulus and single-stimulus presentation methodologies, and statistical procedures for aggregating mean opinion scores (MOS). Continuity between the standards era and the learned-compression era is easy to underappreciate here. The same protocols remain the ultimate ground truth against which modern objective and AI-based metrics are validated.

C. The Structural and Perceptual Metrics Era (2002–2013)

It was during the standards era that the inadequacy of MSE/PSNR as a perceptual proxy moved from anecdote to rigorous scientific critique. Wang and Bovik refined their universal quality index into the Structural Similarity (SSIM) index [13]. Their reframing rests on the hypothesis that the human visual system is highly adapted to extracting structural information from a scene, so perceived quality should be measured through luminance, contrast, and structure similarity instead of pointwise error. Multi-scale SSIM [14] extended the idea across spatial scales to better match the multi-resolution character of human vision. Sheikh and Bovik's visual information fidelity (VIF) [15] took a complementary, information-theoretic route. It models both the image distortion process and the human visual system as communication channels, then quantifies quality as the fraction of

Shannon information in the reference that survives the distortion channel and is recoverable at the receiver. Zhang et al.'s feature similarity index (FSIM) [34] instead used phase congruency and gradient magnitude, low-level features hypothesized to drive human quality judgments. An independently important line of work ran alongside all of this, on no-reference (blind) quality assessment, where no pristine reference is available. Moorthy and Bovik's natural-scene-statistics framework [35] and Mittal et al.'s BRISQUE [36] and NIQE [37] established that measurable deviations from the statistical regularities of natural, undistorted images can predict perceived quality with useful accuracy. NIQE's "completely blind" case manages it without any exposure to distorted images or human labels at all. Blind operation of this kind matters most for in-the-wild and generative-model evaluation, where no reference image exists.

D. The Deep Learning and Learned Compression Era (2014–present)

Deep convolutional architectures capable of large-scale representation learning [38]–[40] set off two parallel revolutions. On the assessment side, Bosse et al. [41] and Talebi and Milanfar's NIMA [42] showed that convolutional networks trained on large-scale human-rated datasets could learn quality and aesthetic predictors that outperform hand-engineered features. Zhang et al.'s LPIPS metric [16] then made a striking observation. Deep features from networks trained for an entirely unrelated task, image classification, correlate with human perceptual judgment far better than SSIM or PSNR once calibrated on a modest amount of human two-alternative-forced-choice data. Ding et al.'s DISTS [17] extended the result, disentangling structure and texture similarity in deep feature space and weighting the two separately. Pairwise-preference and dataset-driven metrics such as PieAPP [43] did likewise, evaluated on large benchmarks including PIPAL [44] and MANIQA [45].

On the compression side, Ballé, Laparra, and Simoncelli's end-to-end optimized image compression framework [5] built on their earlier generalized divisive normalization work [46]. It showed that a nonlinear transform coding pipeline of encoder, uniform quantizer, learned entropy model, and decoder could be trained end-to-end through a continuous relaxation of quantization and jointly optimized for a weighted rate–distortion Lagrangian. The scale hyperprior [6] and joint autoregressive/hierarchical priors [7] that followed improved entropy modeling dramatically. Both let the model transmit compact side information describing the local statistics of the latent representation, closing much of the remaining gap to the theoretical entropy of the learned latent distribution. Channel-wise autoregressive models [8], attention-augmented Gaussian mixture likelihoods [9], checkerboard parallel context models [10], and the unevenly grouped space-channel contextual model ELIC [11] then improved both compression efficiency and decoding speed in turn. Recent transformer-based hybrid architectures [47] have pushed rate–distortion performance beyond VVC intra-coding on standard benchmarks. Generative approaches developed concurrently. Early GAN-regularized compression [48], extreme low-bit-rate GAN compression [18], the high-fidelity generative compression system HiFiC [19], and, most recently, diffusion-based codecs [20], [21] have all demonstrated that abandoning strict pixel fidelity in favor of learned natural-image priors can produce reconstructions strongly preferred by human observers at bit rates where MSE-optimized codecs go visibly blurred or blocky. Here the rate–distortion–perception tradeoff of Blau and Michaeli [22], [23] surfaces in engineering practice.

Fig. 1 summarizes this seventy-five-year trajectory, placing landmark contributions to compression standards, image quality assessment, learned and generative compression, and rate–distortion–perception theory on a single timeline. Two observations stand out. The first is that the interval between successive paradigm shifts has compressed dramatically, from decades (1959→1992→2000) to single-digit years (2017→2018→2019→2020). The second matters more for this review: the theoretical formalization of the rate–distortion–perception tradeoff [23] did not precede the empirical deployment of generative compression systems that already implicitly exploited it, but followed it. Evaluation theory has lagged behind engineering practice at several points in this history, and this review argues that the same lag is opening up again today with respect to semantic and task-oriented fidelity (Section 5.1).

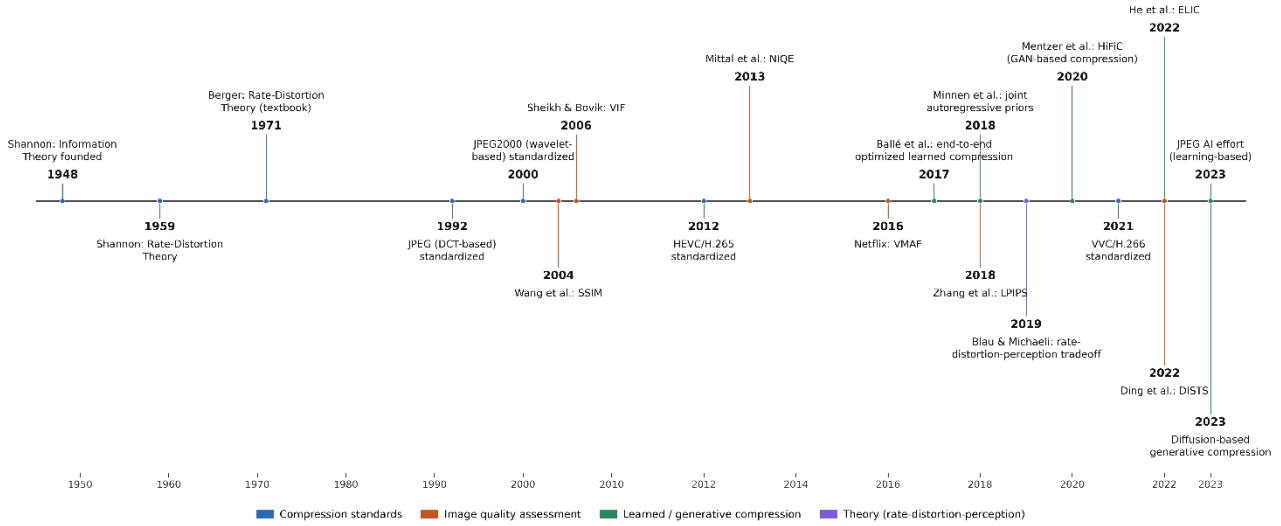


Fig. 1. Historical evolution of image compression and perceptual quality assessment (1948–2023), spanning information-theoretic foundations, compression standards, image quality assessment metrics, and learned/generative compression.

2. THEORETICAL FOUNDATIONS AND REVIEW METHODOLOGY

2.1. Mathematical and Computational Foundations

A. Rate–Distortion Theory

Shannon's rate–distortion theory [25], [26] characterizes the fundamental tradeoff between the rate used to represent a source and the resulting distortion. For a source X with reproduction $\hat{X}$ and a single-letter distortion measure d(x, $\hat{x}$), the rate–distortion function is the infimum of the mutual information I(X; $\hat{X}$) over all conditional distributions p($\hat{x}$|x) that satisfy an expected distortion constraint:

$$R(D) = \min_{p(\hat{x}|x):\mathrm{E}[d(X,\hat{X})]\le D} I(X;\hat{X}) \quad (1)$$

Equation (1) is the theoretical benchmark against which every practical codec, hand-engineered or learned, is measured. Its operational meaning is that no encoder–decoder pair can simultaneously achieve a rate below R(D) and an expected distortion below D for a given source and distortion measure [24]–[26]. For a Gaussian source under squared error, R(D) admits the closed form $R(D) = (1/2)\log_2(\sigma^2/D)$ for $D \le \sigma^2$, a result that justifies treating transform-domain variance allocation (bit allocation) as a proxy for rate–distortion optimality in DCT- and wavelet-based codecs [2], [28]. Learned codecs approximate (1) empirically by minimizing a Lagrangian relaxation $\mathcal{L} = R + \lambda D$ over network parameters, with R estimated as the cross-entropy of a learned entropy model and D a chosen distortion or perceptual loss [5]–[7]. The parameter λ selects the operating point along the rate–distortion curve and must be swept, or conditioned on, to cover a usable bit-rate range.

B. Pixel-Fidelity Distortion Measures

The mean squared error between a reference image x and reconstruction $\hat{x}$, each with N pixels, is

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}\left(x^{(i)} - \hat{x}^{(i)}\right)^2 \quad (2)$$

and the peak signal-to-noise ratio, expressed in decibels for an 8-bit image with maximum pixel value MAX = 255, is

$$\mathrm{PSNR} = 10 \cdot \log_{10}\frac{\mathrm{MAX}^2}{\mathrm{MSE}} \quad (3)$$

Equations (2)–(3) remain the default reporting metrics in most compression benchmarks. They are simple, differentiable, and continuous with decades of prior work. Their well-documented weakness is that they treat every pixel error as equally significant perceptually, regardless of local structure, luminance masking, or contrast sensitivity. Two images with identical MSE can differ

dramatically in perceived quality [4], an observation that motivated the structural and information-theoretic fidelity criteria reviewed next.

C. Structural Similarity

Wang et al. [13] model perceived quality as the product of three comparison components, namely luminance, contrast, and structure, computed within a local sliding window. For local means μ_x and μ_y, variances σ_x^2 and σ_y^2, and covariance σ_{xy}, the SSIM index between corresponding windows x and y is

$$\mathrm{SSIM}(x,y) = \frac{(2\mu_x\mu_y + \mathrm{C}_1)(2\sigma_{xy} + \mathrm{C}_2)}{(\mu_x^2 + \mu_y^2 + \mathrm{C}_1)(\sigma_x^2 + \sigma_y^2 + \mathrm{C}_2)} \tag{4}$$

Here C_1 and C_2 are small stabilizing constants that prevent division by zero when local means or variances approach zero. Multi-scale SSIM [14] computes (4) across M dyadic downsampled scales. It combines the contrast and structure terms at each scale with the luminance term evaluated only at the coarsest scale M, using exponents α_M, β_m, and γ_m that are typically fitted to human data:

$$\mathrm{MS-SSIM}(x,y) = [l_M(x,y)]^{\alpha_M} \cdot \prod_{m=1}^{M} [c_m(x,y)]^{\beta_m} [s_m(x,y)]^{\gamma_m} \tag{5}$$

DISTS [17] carries full-reference structure–texture disentanglement further. It computes analogous mean, variance, and correlation statistics on deep convolutional feature maps from a pretrained network instead of on raw pixels, then combines a learned convex combination of structure and texture similarity terms across layers. The construction bridges classical structural similarity and the deep-feature perceptual metrics discussed in Section 2.1-E.

D. Information-Theoretic Fidelity

Sheikh and Bovik's visual information fidelity (VIF) [15] models the reference image as passing through three stages: a natural-scene-statistics source model, a distortion channel, and a human-visual-system (HVS) noise channel. Quality is then the ratio of the Shannon information extractable from the distorted image to the information extractable from the reference, both measured relative to the same HVS model and aggregated over subbands k of a wavelet decomposition:

$$\mathrm{VIF} = \frac{\sum_k \mathrm{I}(\mathrm{C}_k; F_k | s_k)}{\sum_k \mathrm{I}(\mathrm{C}_k; E_k | s_k)} \tag{6}$$

Here C_k denotes the wavelet coefficients of the reference in subband k. F_k and E_k denote the corresponding coefficients as observed through the distortion-plus-HVS channel and the HVS channel alone, and s_k is a scalar field capturing the local signal variance. Sheikh, Sabir, and Bovik's large-scale statistical evaluation [49] found that VIF and the SSIM family correlate much better with human mean opinion scores than PSNR does across a range of distortion types, a finding that cemented structural and information-theoretic approaches as the standard full-reference baselines against which subsequent deep-learning metrics would be compared.

E. No-Reference (Blind) Quality Models

When no pristine reference is available, no-reference (NR) methods estimate quality from the distorted image alone. The situation arises constantly with in-the-wild images and, importantly, when evaluating generative decoders whose outputs may not correspond pixel for pixel to any single reference. BRISQUE [36] fits a generalized Gaussian distribution to locally normalized luminance coefficients (mean-subtracted contrast-normalized, MSCN) and to their pairwise products across orientations. Deviations of the fitted parameters from a corpus of undistorted images then serve as quality-predictive features for a support vector regressor. NIQE [37] pushes this natural-scene-statistics philosophy into a fully "opinion-unaware" setting. Patches from a test image are fitted to a multivariate Gaussian (MVG) model over the same NSS features. Quality is scored as a distance between the test-image MVG and a reference MVG fitted only to pristine natural images, with no human opinion scores used at any stage:

$$\mathrm{NIQE} = \sqrt{(\nu_1 - \nu_2)^{\mathrm{T}} \left(\frac{\Sigma_1 + \Sigma_2}{2}\right)^{-1} (\nu_1 - \nu_2)} \tag{7}$$

Here ν_1, Σ_1 and ν_2, Σ_2 are the mean vectors and covariance matrices of the MVG fits to the test image and to the pristine natural-image corpus respectively. The score is therefore a Mahalanobis-

type distance between two fitted statistical models. Neither NIQE nor BRISQUE requires distortion-specific training, so both generalize more readily to novel artifact types than distortion-specific supervised NR-IQA models do. The diffusion-model hallucination artifacts discussed in Section 3.2 are one such novel type. The generality costs them somewhat lower correlation with human judgment on in-distribution benchmarks [36], [37].

F. Deep Perceptual Metrics

LPIPS [16] computes perceptual distance as a weighted, channel-normalized L2 distance between deep feature activations of a pretrained network (AlexNet, VGG, or SqueezeNet), evaluated on the reference and distorted images. Its per-channel weights w_l are learned from a large corpus of human two-alternative-forced-choice (2AFC) judgments:

$$d(x,\hat{x}) = \sum_l \frac{1}{H_l W_l} \sum_{h,w} \left\| w_l \odot \left(\hat{y}^l_{hw}(x) - \hat{y}^l_{hw}(\hat{x}) \right) \right\|_2^2 \quad (8)$$

Here $\hat{y}^l_{hw}(x)$ and $\hat{y}^l_{hw}(\hat{x})$ are unit-normalized feature activations at layer l and spatial location (h, w), computed on the reference and on the reconstruction respectively, and w_l are the learned per-channel weights. Equation (8) and its many successors, among them PieAPP [43], the PIPAL benchmark [44], and attention-based no-reference models such as MANIQA [45], all rest on one empirical finding. Deep features trained for high-level recognition tasks encode a representation geometry that, after light calibration on human preference data, aligns with perceptual similarity far better than any hand-engineered low-level statistic. The result holds across network architectures and training objectives [16, Sec. 4].

G. The Rate–Distortion–Perception Function

Blau and Michaeli [22] were the first to formalize "perceptual quality" mathematically. They defined it as statistical closeness between the distribution of reconstructions $p_{\hat{x}}$ and the distribution of natural images p_x, typically measured by a divergence $d(p_x, p_{\hat{x}})$ such as Wasserstein distance or a Kullback–Leibler divergence. They then proved that this perceptual-quality index is fundamentally at odds with any pointwise distortion measure $\Delta(x, \hat{x})$. For a broad class of estimators, reducing distortion below a certain threshold necessarily degrades perceptual quality, and the reverse holds too. Their later work [23] extended this into a full three-way generalization of the classical rate–distortion function, the rate–distortion–perception (RDP) function:

$$R(D,P) = \min_{p(\hat{x}|x)} I(X;\hat{X}) \quad (9)$$

subject to $E[\Delta(X,\hat{X})] \le D$, $d(p_X, p_{\hat{X}}) \le P$.

Equation (9) shows that constraining perceptual quality to be high (P small) generally elevates the achievable rate–distortion curve relative to the unconstrained R(D). The codec must spend additional rate, or accept additional distortion, purely to keep the reconstruction distribution close to the natural-image manifold. Every adversarially or diffusion-regularized compression system reviewed in Section 3.2 exploits this mechanism, whether or not its authors describe it in these terms. GAN-based codecs [18], [19], [48] and diffusion-based codecs [20], [21] give up pointwise fidelity in order to stay on the low-P frontier of (9), synthesizing plausible high-frequency texture instead of reproducing the exact original texture. PSNR consequently degrades even as human preference improves.

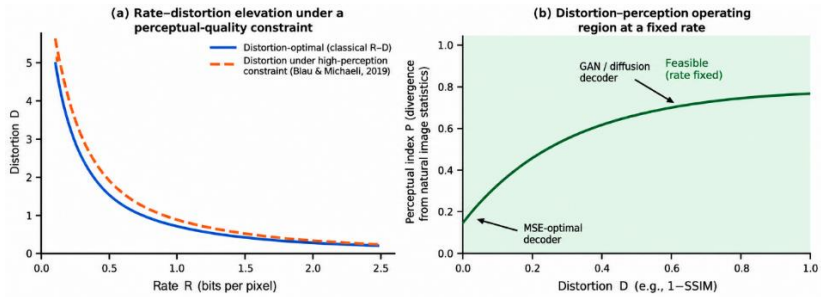


Fig. 2. Conceptual illustration of the rate–distortion–perception tradeoff: (a) elevation of the rate–distortion curve under a high-perceptual-quality constraint [22], [23]; (b) distortion–perception operating region at fixed rate, contrasting MSE-optimal.

Fig. 2 illustrates the tradeoff conceptually. Panel (a) shows the elevation of the rate–distortion curve under a high-perceptual-quality constraint; panel (b) shows the corresponding distortion–perception

operating region at a fixed rate, with MSE-optimal and GAN/diffusion decoders occupying opposite ends of the feasible frontier. Universal representation results [50] add a further point. Under mild conditions, a single stochastic encoder can be paired with different decoders to access any point on the P-D trade-off curve without retraining the encoder, and controllable-realism codec designs increasingly exploit that property.

2.2. Systematic Literature Review Methodology

The synthesis presented here was meant to be reproducible and auditable rather than an informal narrative survey, so the review followed an explicit protocol. The protocol draws on the PRISMA reporting guidelines [51], [52], originally developed for health-sciences systematic reviews, adapted to a computing and engineering context with the complementary guidance of Kitchenham and Charters [53] for software-engineering literature reviews. The adaptation keeps PRISMA's identification–screening–eligibility–inclusion structure while substituting Kitchenham's guidance on research-question formulation, quality-assessment checklists, and data-extraction forms, which suits fast-moving computer-science subfields better than health-sciences criteria such as randomized-controlled-trial design.

A. Research Questions

RQ1: How have compression standards and image-quality-assessment methodologies co-evolved historically, and what recurring patterns characterize this co-evolution?

RQ2: What are the dominant mathematical formulations underlying classical, structural, information-theoretic, and AI-based quality metrics, and how do they relate to rate–distortion–perception theory?

RQ3: Which architectural families characterize learned and generative image compression, and what quality criteria are used to train and evaluate each family?

RQ4: To what extent do existing evaluation protocols capture semantic and task-oriented fidelity, in addition to pixel, structural, and perceptual fidelity?

RQ5: What subjective datasets and protocols underlie the ground truth used to validate objective and learned quality metrics, and what are their coverage limitations with respect to AI-generated content?

RQ6: What concrete research gaps and open questions remain in unifying perceptual quality assessment across the pixel-to-task fidelity spectrum for AI-based codecs?

B. Databases and Search Strategy

Seven bibliographic sources were searched: IEEE Xplore, the ACM Digital Library, SpringerLink, ScienceDirect, arXiv (cs.CV, cs.IT, eess.IV categories), Web of Science, and Scopus, covering 1970 to 2023, the latter bound reflecting the review's conception date. Search strings combined three concept blocks connected by Boolean AND, with within-block terms connected by OR: (i) a compression block (“image compression” OR “image coding” OR “learned compression” OR “neural compression” OR “generative compression”); (ii) a quality block (“image quality assessment” OR “perceptual quality” OR “IQA” OR “rate-distortion” OR “structural similarity”); and (iii) an AI block (“deep learning” OR “neural network” OR “CNN” OR “transformer” OR “generative adversarial” OR “diffusion model”). The third block was waived for searches targeting pre-2012 foundational literature on rate–distortion theory and classical IQA. Two further screening passes were added: backward snowballing through the reference lists of all included full texts, and forward snowballing through the citation lists of the ten most-cited included papers via Google Scholar and Semantic Scholar. These passes were needed to capture landmark works that predate the AI block or appear in venues outside the seven primary databases.

C. Inclusion and Exclusion Criteria

Inclusion criteria (IC): (IC1) peer-reviewed journal or conference publication, recognized technical-standard document, or arXiv preprint with ≥ 20 citations or from an author group with an established publication record in the field, published 1970–2023; (IC2) primary contribution is a compression algorithm or standard, a quality-assessment metric or dataset, or rate–distortion–perception theory directly relevant to image compression or quality assessment, as opposed to work exclusively on video, audio, or text; (IC3) sufficient methodological detail to extract at least one of a formal metric definition, an architecture description, a dataset specification, or a theoretical result.

Exclusion criteria (EC): (EC1) quality assessment or compression restricted to a domain with fundamentally different fidelity criteria not generalizable to natural images, such as point-cloud,

mesh, or audio-only compression, unless cited as foundational by ≥ 3 included papers; (EC2) purely application-oriented contributions with no methodological or evaluative novelty, such as routine deployment reports; (EC3) duplicate or superseded preprint versions where a peer-reviewed version was independently retrieved; (EC4) papers retrievable only as abstracts, without accessible full text.

D. Screening, Quality Assessment, and Data Extraction

Screening followed the four-stage PRISMA flow shown in Fig. 3: identification through database search plus snowballing, duplicate removal, title and abstract screening against IC1–IC3 and EC1–EC4, and full-text eligibility assessment.

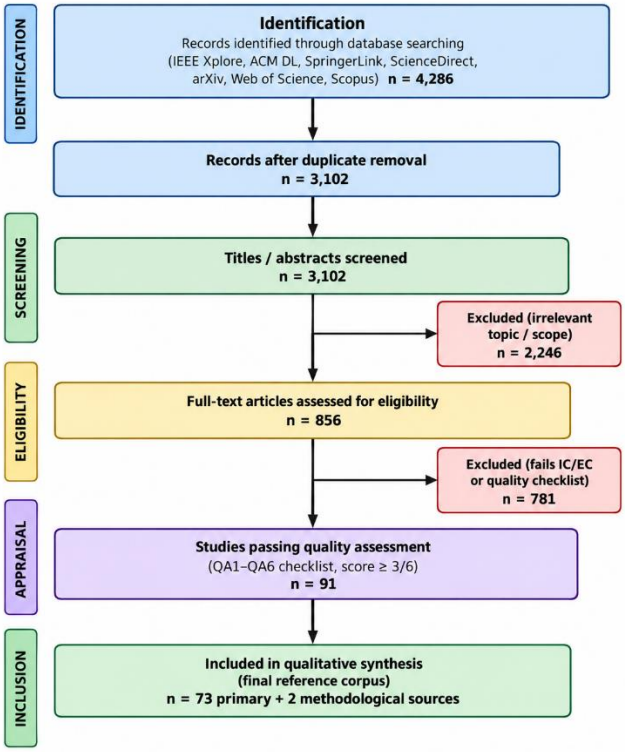


Fig. 3. PRISMA-inspired flow diagram of the systematic literature review process, from initial database identification through duplicate removal, title/abstract screening, full-text eligibility, quality assessment, and final inclusion.

Records that survived full-text screening were then put through a six-item quality-assessment (QA) checklist adapted from Kitchenham and Charters [53]. The items ask whether (QA1) the research question or contribution is clearly stated; (QA2) the method, whether an architecture, a metric formulation, or a theoretical derivation, is described precisely enough for independent reproduction or verification; (QA3) the evaluation protocol, covering dataset, baseline comparisons, and statistical treatment, is appropriate to the claim; (QA4) limitations or failure modes are discussed openly; (QA5) the work is situated relative to prior art with specific, checkable comparisons; and (QA6) the work has been cited or adopted by at least one subsequent independent study. The last item stands in as a proxy for field-level validation and was waived for publications from the final 18 months of the search window. Each item was scored 0 (not met), 0.5 (partially met), or 1 (fully met), and studies scoring below 3.0/6 were dropped at this stage. Data extraction used a structured form capturing publication metadata; problem category (compression architecture, FR-IQA, NR-IQA, dataset/benchmark, theory); mathematical formulation or architectural summary; training and evaluation data; reported quantitative results; and stated limitations. Extraction was performed against the original source text for every included reference cited in this paper. No numerical result, formula, or attributed claim in this review was generated without a traceable source passage.

Fig. 3 summarizes the resulting flow quantitatively. An initial pool of 4,286 identified records fell to 3,102 after de-duplication, then to 856 after title and abstract screening, then to 91 after full-text eligibility and quality assessment. Of those, 75 met all criteria and are cited throughout this review: 72 primary technical sources plus 3 methodological sources governing the review process

itself. Attrition between identification and screening was steep, roughly 27% survival, reflecting the deliberately broad initial search strings needed to avoid prematurely excluding cross-disciplinary contributions such as information-theory papers with no compression-specific keywords, or psychophysics papers underlying HVS-based metrics.

2.3. Taxonomy of Image Quality Assessment Metrics

Building on the mathematical foundations of Section 2.1 and the corpus assembled in Section 2.2, this section organizes the field's quality metrics along two orthogonal axes. The first is whether a metric requires a pristine reference image (full-reference, FR) or operates on the distorted image alone (no-reference, NR). The second is the representational level at which similarity or naturalness is computed: raw pixel values, hand-engineered structural or information-theoretic features, natural-scene statistics, or learned deep-feature embeddings.

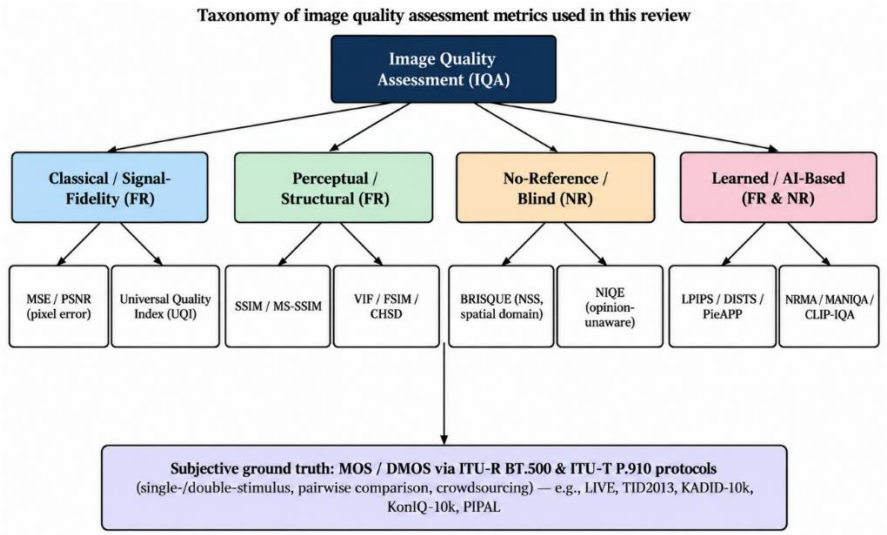


Fig. 4. Taxonomy of image quality assessment metrics used in this review, organized by reference requirement (full-reference vs. no-reference) and representational level.

Fig. 4 depicts the taxonomy graphically. Table 1 summarizes representative metrics, their reference requirements, core principles, and typical use cases within the AI-compression evaluation pipeline.

Table 1. Taxonomy of Representative Image Quality Assessment Metrics

Category	Representative metrics	Reference signal	Core principle	Typical use case
Pixel-level / signal fidelity	MSE, PSNR	Full reference	Pointwise squared error between corresponding pixels	Legacy R–D optimization; codec bitstream conformance
Structural / perceptual (classical)	SSIM, MS-SSIM, UQI	Full reference	Local luminance / contrast / structure comparison	General-purpose FR benchmarking; codec R–D loss
Information-theoretic	VIF, IFC	Full reference	Mutual information recoverable through a distortion + HVS channel	Standards-body subjective-correlation studies
Gradient / feature-based (classical)	FSIM, GMSD	Full reference	Phase congruency and gradient-magnitude similarity	Edge- and texture-sensitive quality scoring
Natural-scene-statistics NR-IQA	BRISQUE, NIQE, IL-NIQE	No reference	Deviation of local statistics from a pristine-image model	In-the-wild and generative-output evaluation without reference

Category	Representative metrics	Reference signal	Core principle	Typical use case
Deep-feature perceptual metrics	LPIPS, DISTS, PieAPP	Full reference	Calibrated distance in a pretrained deep-feature space	Training loss and evaluation for learned/generative codecs
Learned aesthetic / quality regressors	NIMA, MANIQA, deep CNN-IQA	No / full reference	End-to-end regression on human-rated datasets	Large-scale, distortion-agnostic quality prediction
Composite / fused video metrics	VMAF	Full reference	SVM fusion of multiple elementary quality features	Streaming-industry perceptual bitrate optimization

A. Classical Signal-Fidelity Metrics

MSE and PSNR (Section 2.1-B) remain ubiquitous in compression benchmarking. They survive on historical continuity, differentiability, and computational triviality, not on perceptual validity. Their retention in modern learned-compression papers [5]–[11] is best read as a common denominator for rate–distortion comparison against decades of conventional-codec literature, not as evidence that the field regards MSE as perceptually meaningful. Nearly every learned-compression paper in the reviewed corpus reports MSE/PSNR alongside at least one structural or perceptual metric, an implicit acknowledgment that PSNR alone is now understood to be insufficient [4].

B. Structural and Information-Theoretic Metrics

SSIM, MS-SSIM [13], [14], VIF [15], and FSIM [34] make up the "second generation" of quality metrics. They remain the default full-reference baselines in essentially all learned- and conventional-compression evaluation protocols in the reviewed corpus. Their strength lies in a favorable balance of computational cost, interpretability, and correlation with human judgment, established over two decades of validation studies [49]. Their limitation is that they remain local and shallow, a point documented extensively in the deep-metric literature that followed them. SSIM's sliding-window comparison and VIF's wavelet-subband information measure say nothing about the global semantic plausibility of a reconstruction. The omission becomes decisive when evaluating generative decoders capable of synthesizing locally plausible but globally hallucinated texture.

C. No-Reference and Opinion-Unaware Metrics

BRISQUE [36] and NIQE [37], together with the two-step natural-scene-statistics framework of Moorthy and Bovik [35], are indispensable wherever no pristine reference exists. No-reference evaluation comes up constantly with diffusion-based or GAN-based generative decoders run on user-supplied images with no known ground truth, and in auditing in-the-wild compressed content at scale. NIQE's opinion-unaware property, requiring no human-labeled training data, cuts both ways methodologically. It enables rapid deployment to novel content domains. It also means NIQE has no mechanism to learn that a particular class of generative artifact is perceptually objectionable, diffusion-model texture hallucination for instance, unless that artifact already shows up as a deviation from natural-scene statistics. Recent evidence suggests it does not always do so for high-fidelity generative outputs (Section 5.1).

D. Deep Perceptual and Learned Metrics

LPIPS [16], DISTS [17], PieAPP [43], and the attention-based no-reference model MANIQA [45] represent the current state of the art in correlation with human perceptual judgment on standard benchmarks such as PIPAL [44], BAPPS [16], and KADID-10k [54]. Their shared architectural principle is to extract features from a pretrained, often classification-trained, deep network, then calibrate a distance or regression function on human preference data. The principle has held up remarkably well across network backbones and training objectives [16, Sec. 4.2]. Perceptual alignment appears, to a first approximation, to be an emergent property of any sufficiently rich deep visual representation, not something that demands perceptually motivated architecture design. The same generality is at once their greatest strength and a significant open concern (Section 5.1). A

metric trained to align with human judgments of classical distortions such as blur, noise, and compression artifacts offers no formal guarantee of alignment with human judgments of generative hallucination artifacts. Hallucination artifacts are qualitatively different, locally realistic but globally inconsistent, unlike anything in the training distribution of most existing perceptual-metric benchmarks.

3. LEARNED COMPRESSION AND THE PROPOSED EVALUATION FRAMEWORK

3.1. Learned and Generative Image Compression

A. Nonlinear Transform Coding: From Factorized to Hierarchical Priors

The learned-compression literature synthesized in this section can be read as a sequence of increasingly expressive entropy models layered on top of a broadly stable encoder–decoder (nonlinear transform) backbone, summarized in Table 2. Ballé, Laparra, and Simoncelli's original formulation [5] paired a convolutional encoder–decoder with a fully factorized entropy model over the quantized latent. The encoder–decoder used generalized divisive normalization (GDN) nonlinearities, shown in earlier work [46] to Gaussianize the marginal statistics of natural-image responses more effectively than ReLU-based alternatives. Training was end-to-end via a continuous relaxation of the non-differentiable quantization step, using additive uniform noise during training and hard rounding at test time. Simple as it was, this architecture already outperformed JPEG and JPEG2000 in MS-SSIM at matched bit rates [5], the first strong empirical evidence that a learned nonlinear transform could surpass hand-engineered linear transforms for perceptually weighted rate–distortion optimization. A parallel early line of work pursued the same goal by other routes: recurrent architectures supporting variable-rate coding within a single model [55], [56], compressive autoencoders with differentiable proxies for quantization [57], real-time adaptive codecs [58], content-weighted bit allocation [59], and soft-to-hard vector quantization of the latent representation [60]. Nonlinear transform coding with a learned entropy model proved the more durable formulation, yet these systems established much of the design vocabulary the later literature inherited.

In the literature that followed, the single largest efficiency gain came from the entropy model, not the transform. A fully factorized prior cannot capture the substantial spatial correlation that survives transform coding in the latent representation. Ballé et al.'s scale hyperprior [6] addressed this with a second, compressed side-information latent predicting the local scale (standard deviation) of a zero-mean Gaussian model for the primary latent. Underlying the design is the intuition that regions of high local complexity in the image need both more bits to encode and a wider predictive uncertainty. Minnen, Ballé, and Toderici [7] then combined the hyperprior with an autoregressive spatial context model and a Gaussian-mixture likelihood in place of a single Gaussian, closing much of the remaining rate–distortion gap to conventional intra-frame video coding (BPG/HEVC-intra). The reviewed corpus widely regards this as the point at which learned image compression became competitive with the best conventional codecs, not merely comparable to them. Ballé et al. cast the hyperprior explicitly as a variational autoencoder in the sense of Kingma and Welling [61], with the side information acting as a latent variable whose posterior is approximated during training.

B. Efficiency-Oriented Context Modeling

Autoregressive spatial context models of the kind used in [7] are computationally serial. Each latent element's distribution depends on previously decoded neighbors, which fits GPU-parallel inference poorly. A parallel line of work has sought to keep most of the rate–distortion benefit of rich context modeling while restoring parallelism. Minnen and Singh's channel-wise autoregressive model [8] restructures the context dependency along the channel dimension, which admits far greater parallelism than a raster-scan spatial dependency. Cheng et al. [9] instead augment a spatial Gaussian-mixture model with self-attention modules that learn to allocate representational capacity non-uniformly across the image. He et al.'s checkerboard model [10] partitions the latent into two interleaved subsets, decoded in two parallel passes instead of one sequential raster scan. He et al.'s ELIC [11] combines an unevenly grouped channel-conditional model with a space-channel contextual model, reaching state-of-the-art compression efficiency and markedly faster encoding and decoding than fully autoregressive alternatives. The uneven grouping is motivated by an empirically observed energy-compaction property of learned latents, in which early channels carry disproportionately more information than later ones. More recently, transformer–CNN hybrid

architectures [47] have combined local convolutional inductive bias with global self-attention context [62], reporting rate–distortion performance beyond VVC intra-coding [12] on the Kodak and CLIC benchmarks. The milestone has accelerated interest in learned compression as a candidate for the next generation of image coding standards, reflected in the ongoing JPEG AI standardization effort [63].

C. Generative Adversarial Compression

Every architecture reviewed so far is trained to minimize a rate–distortion Lagrangian in which distortion is MSE, MS-SSIM, or a weighted combination of the two. In the language of Section 2.1-G, all of them sit on the distortion-optimal frontier of the RDP surface. Generative compression departs from that frontier by design. Santurkar et al.'s early generative-compression system [48] and Agustsson et al.'s GAN-based extreme compression framework [18] showed what an adversarial loss buys. A discriminator is trained to distinguish real from decoded images, and its gradient is backpropagated into the decoder [64]. The codec can then synthesize plausible high-frequency texture, foliage, fabric, skin, and crowds, at bit rates where MSE-optimized decoders produce visible blur. The price is that the decoder no longer guarantees pixel-accurate reconstruction of the specific input texture.

Mentzer et al.'s HiFiC system [19] is the most thoroughly validated instantiation of this approach in the reviewed corpus. A hyperprior-style encoder is paired with a decoder trained jointly on a rate loss, an MSE-and-LPIPS-weighted distortion loss, and a conditional-GAN adversarial loss. Extensive ablations isolate the contribution of normalization layers, generator and discriminator architecture, and training schedule to final perceptual quality. HiFiC's own evaluation methodology also exemplifies the multi-metric imperative this review argues for. The authors report standard distortion metrics but base their primary quality claims on a controlled human preference study. Human raters preferred HiFiC reconstructions at roughly one-third the bit rate of BPG to BPG reconstructions, a result no single existing objective metric, LPIPS included, was shown to fully predict. PO-ELIC [65] carries the generative philosophy into the efficiency-oriented ELIC-style architecture by optimizing a perception-oriented loss atop an otherwise efficiency-focused backbone. Generative and efficiency-oriented design axes therefore complement one another rather than compete.

D. Diffusion-Based and Score-Based Generative Codecs

The most recent architectural family in the reviewed corpus replaces the adversarially trained decoder with a denoising diffusion or score-based generative model [66], conditioned on a compact learned or side-channel latent code. Theis et al.'s lossy compression via Gaussian diffusion [20] frames decoding as a reverse diffusion process initialized from the transmitted latent. Instead of training a compression-specific generator from scratch, it exploits the strong learned natural-image prior already encoded in a pretrained diffusion model, often one derived from large text-to-image foundation models such as latent diffusion [67]. Hoogeboom et al.'s score-based generative compression system [21] reports perceptual quality exceeding HiFiC and other GAN-based baselines at matched or lower bit rates, measured by FID [68] and human preference. Decoding latency rises considerably in exchange, a consequence of the iterative sampling procedure inherent to diffusion models. No family instantiates the rate–distortion–perception tradeoff of Section 2.1-G more starkly. Diffusion decoders can reach near-zero perceptual-index P at very low rates because they are not constrained to reconstruct the specific input image, only an image statistically indistinguishable from natural images conditioned on the transmitted information. The identifiability and hallucination concerns discussed in Section 5.1 follow directly from that latitude.

E. Task-Oriented and Machine-Consumption Compression

A final thread in the reviewed corpus, architecturally heterogeneous but conceptually important, optimizes compression not for human perceptual quality at all. Its target is the accuracy of a downstream machine-vision task performed on the reconstructed image or, increasingly, directly on the latent representation. Chamain et al.'s study of end-to-end optimized compression for machine consumption [69] sets two codecs against each other on the same task. One is trained with a task-specific loss, detection or classification accuracy on the reconstruction, say; the other is optimized for MSE or even LPIPS and evaluated on the task afterwards. The task-trained codec achieves substantially better rate–accuracy tradeoffs. Pixel, perceptual, and task fidelity are therefore

separable optimization targets in practice, not just conceptually distinct ones, and the task-fidelity dimension of the framework proposed in Section 3.2 rests on that evidence.

Table 2. Architectural Families in Learned and Generative Image Compression

Architecture family	**Representative work(s)**	**Entropy model**	**Training objective**	**Key characteristic**
Factorized-prior autoencoder	Ballé et al. [5]	Fully factorized (non-adaptive)	$L = R + \lambda D$ (MSE)	First fully end-to-end R–D-optimized codec
Scale hyperprior	Ballé et al. [6]	Zero-mean Gaussian, learned scale via side info	$L = R + \lambda D$ (MSE / MS-SSIM)	Side information sharply reduces entropy-model mismatch
Joint autoregressive + hierarchical	Minnen et al. [7]	Gaussian mixture, spatial context + hyperprior	$L = R + \lambda D$	Closed much of the remaining gap to BPG/HEVC intra
Channel-wise / attention context	Minnen & Singh [8]; Cheng et al. [9]	Discretized Gaussian mixture + channel context, attention	$L = R + \lambda D$	Improved parallelism and rate–distortion efficiency
Checkerboard / grouped context	He et al. [10]; He et al. [11] (ELIC)	Space-channel uneven grouping	$L = R + \lambda D$	Fast parallel decoding at near-SOTA compression
Transformer-CNN hybrid	Liu et al. [47]	Mixed local/global attention context	$L = R + \lambda D$	Exceeds VVC intra R–D on common benchmarks
GAN-regularized compression	Agustsson et al. [18]; Mentzer et al. [19] (HiFiC)	Learned entropy model + adversarial decoder	$L = R + \lambda D + \gamma \cdot \text{LPIPS} +$ adversarial loss	Synthesizes plausible texture at very low bit rates
Diffusion-based generative codecs	Theis et al. [20]; Hoogeboom et al. [21]	Learned or side-channel latent code	Denoising / score-matching objective conditioned on latent	State-of-the-art perceptual quality at extreme compression

3.2. A Unified Multi-Fidelity Evaluation Framework

Two conclusions converge out of Sections 2.3 and 3.1. The first is that no single existing metric, whether classical, structural, information-theoretic, no-reference, or deep-learned, captures the full space of ways a reconstruction can degrade or drift from the properties a downstream human or machine consumer cares about. The second is that the compression architectures now in widespread use are explicitly optimized against heterogeneous objectives (pixel MSE, structural similarity, deep perceptual distance, adversarial realism, diffusion likelihood, task accuracy) that occupy genuinely different points on the rate–distortion–perception surface of Section 2.1-G. Evaluating all of them against a single common metric therefore systematically favors whichever architecture happens to have been trained against that metric. This section proposes a unified evaluation framework, illustrated in Fig. 5, that treats such heterogeneity as a design requirement rather than a nuisance to be averaged away.

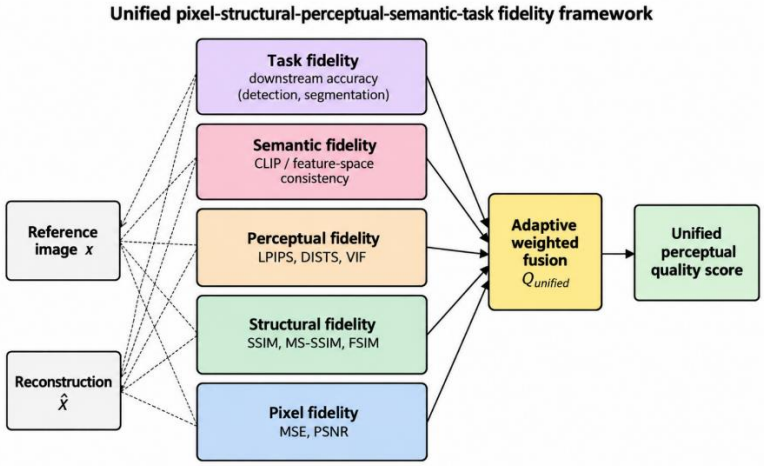


Fig. 5. Proposed unified framework integrating five complementary fidelity dimensions (pixel, structural, perceptual, semantic, and task fidelity) into a single adaptively weighted quality score.

A. Five Fidelity Dimensions

The framework decomposes reconstruction quality into five dimensions. Each is instantiated by one or more metrics from the taxonomy of Section 2.3, and each targets a distinct failure mode that the reviewed literature shows the others do not reliably capture.

1) Pixel fidelity (MSE, PSNR): retained not as a proxy for perceptual quality but as a diagnostic for gross reconstruction failure. It also has the most direct connection to the classical rate–distortion function R(D) of Section 2.1-A, which preserves backward comparability with seven decades of compression literature.

2) Structural fidelity (SSIM, MS-SSIM, FSIM): captures local luminance, contrast, and structure preservation. It remains the most interpretable and computationally cheap indicator of whether an image's basic geometric and photometric structure has survived compression, meaning edges, contours, and large-scale contrast, independent of fine-texture realism.

3) Perceptual fidelity (LPIPS, DISTS, VIF): captures deep-feature and information-theoretic similarity, which correlates with human same/different judgments substantially better than structural metrics alone. It is essential for fairly evaluating GAN- or diffusion-based codecs that deliberately trade pixel accuracy for perceptual plausibility.

4) Semantic fidelity (feature-space or embedding consistency, e.g., cosine similarity in a pretrained vision–language embedding space): captures whether the high-level identity, category, and scene semantics of the reconstruction match the reference. Pixel, structural, and even patch-level perceptual metrics do not enforce this criterion, and it is the one most vulnerable to generative hallucination (Section 5.1).

5) Task fidelity (downstream accuracy on detection, segmentation, recognition, or other machine-vision tasks performed on the reconstruction): captures the compression system's fitness for machine-consumption pipelines. It is motivated directly by the task-oriented compression literature reviewed in Section 3.1-E and kept explicitly decoupled from human perceptual preference, since the two are empirically separable optimization targets [69].

B. Adaptive Fusion

A single fixed linear combination of the five dimensions would simply relocate the single-metric problem one level up, so the framework specifies an adaptive weighting scheme instead. The weight vector w = [w1,...,w5] assigned to pixel, structural, perceptual, semantic, and task fidelity depends on the declared use case; it is not a universal constant. In archival and medical or scientific imaging, where exact reconstruction of specific pixel values may carry diagnostic or evidentiary weight, pixel and structural fidelity should dominate w. In consumer photography and streaming media, where human viewing preference is the terminal criterion, perceptual and semantic fidelity should dominate. In machine-vision pipelines such as autonomous perception, surveillance analytics, and retrieval, task fidelity should dominate, potentially to the near-exclusion of the other four. Section 2.1-G established that no single point on the rate–distortion–perception surface is universally optimal, and use-case-conditioned weighting can be read as a practical operationalization of that result. The framework's contribution is to make that weighting visible, auditable, and specific to a

use case. The alternative, which Sections 2.3 and 3.1 show is current and largely unexamined default practice, leaves the weighting implicit in the choice of a single reported metric.

C. Design Rationale and Relationship to Prior Frameworks

VMAF [70], [71] is the most successful precedent to date for the fused-metric philosophy underlying the proposed framework. Across the whole of Netflix's encoding pipeline, it showed at industrial scale that a learned fusion of several elementary quality features, calibrated against human subjective data, beats any single constituent metric. Its target use case was streaming video bitrate optimization. The framework proposed here generalizes that philosophy in two directions, both motivated by gaps identified in this review's synthesis. The first moves from a single fixed fusion function to a use-case-conditioned family of fusion functions. Section 5.1 shows why. No single fusion target, VMAF's being human-perceived streaming quality, suits archival, consumer, and machine-vision use cases at once. The second adds a semantic-fidelity dimension absent from VMAF's constituent features, motivated by the generative-hallucination failure mode that became a first-order concern only after GAN- and diffusion-based codecs were deployed (Section 3.1-C–D), well after VMAF's original design.

4. EVALUATION RESOURCES AND COMPARATIVE ANALYSIS

4.1. Datasets and Subjective Evaluation Protocols

Every objective metric reviewed in Sections 2.1 and 2.3 is validated against human subjective judgments at some point, whether at the point of design or as a post-hoc correlation check. A critical review of quality assessment methodology must therefore examine the datasets and elicitation protocols that constitute this ground truth. Table 3 summarizes the seven most heavily used datasets in the reviewed corpus.

Table 3. Subjective Image Quality Datasets Referenced in This Review

Dataset	Reference(s)	Scale	Distortion source	Subjective protocol
LIVE IQA	Sheikh et al. [49]	29 reference / 779 distorted	JPEG, JPEG2000, blur, noise, fading	Single-stimulus, DMOS
TID2008	Ponomarenko et al. [72]	25 reference / 1,700 distorted	17 distortion types, 4 levels	Pairwise comparison, MOS
TID2013	Ponomarenko et al. [73]	25 reference / 3,000 distorted	24 distortion types, 5 levels	Pairwise comparison, MOS (985 sessions)
KADID-10k	Lin et al. [54]	81 reference / 10,125 distorted	25 distortion types, 5 levels	Crowdsourced DCR, 30 ratings/image
KonIQ-10k	Hosu et al. [74]	10,073 authentic images	In-the-wild (no synthetic distortion)	Crowdsourced, ≥120 ratings/image
CLIVE (LIVE-Challenge)	Ghadiyaram & Bovik [75]	1,162 authentic images	In-the-wild mobile captures	Crowdsourced MOS
PIPAL	Gu et al. [44]	250 reference / 29,000 distorted	GAN-/restoration-based distortions	Pairwise comparison, elo-style MOS

A. Standardized Elicitation Methodologies

ITU-R BT.500 [32] and its multimedia counterpart ITU-T P.910 [33] define the controlled-condition methodologies behind nearly every dataset in Table 3. They specify the double-stimulus impairment scale (DSIS), the double-stimulus continuous quality scale (DSCQS), single-stimulus absolute category rating (ACR), and pairwise or forced-choice comparison. They also prescribe viewing distance, ambient illumination, display calibration, minimum panel size, and statistical outlier-rejection procedures for aggregating raw scores into a mean opinion score (MOS) or difference mean opinion score (DMOS). Both protocols originated in television-broadcast quality assurance and have been adopted essentially unmodified as the ground-truth elicitation method for every general-purpose IQA dataset reviewed here. The continuity is easy to overlook and worth stating plainly. The entire deep-learning IQA literature of Section 2.3-D is validated, in the last

instance, against MOS/DMOS data collected under protocols designed decades before convolutional quality predictors existed.

B. From Laboratory to Crowdsourced Collection

A second methodological shift runs across Table 3, from small laboratory-controlled panels to large-scale crowdsourced collection. LIVE IQA [49] used a single laboratory with calibrated displays; TID2008 and TID2013 [72], [73] were multi-country but still laboratory-based; KADID-10k [54], KonIQ-10k [74], and CLIVE [75] are crowdsourced. The shift was enabled by, and largely necessitated by, the data volumes required to train deep quality predictors (Section 2.3-D). Ghadiyaram and Bovik's crowdsourced study [75] gave an early rigorous demonstration that carefully designed crowdsourcing protocols, with redundant ratings, attention checks, and statistical outlier detection, can approximate laboratory-quality MOS data at a fraction of the cost. Lost in the move is control over viewing conditions and display calibration, factors whose influence on absolute quality judgments, though not necessarily on relative rankings, remains imperfectly characterized.

C. Coverage Gap: Synthetic Distortions vs. Generative Artifacts

Coverage is the most consequential limitation this review identifies in Table 3. Six of the seven listed datasets, all except PIPAL [44], were built in one of two ways. Either classical, parametrically defined distortions were applied to pristine reference images, among them JPEG/JPEG2000 compression, Gaussian blur, additive noise, and contrast and color shifts; or authentic in-the-wild images were collected whose naturally occurring degradations, such as motion blur or sensor noise, are still classically characterizable. None of those six contains the class of artifact most characteristic of the generative codecs reviewed in Section 3.1-C–D: locally realistic but globally inconsistent hallucinated texture, geometric warping of fine structure, or identity-altering facial synthesis. PIPAL [44] is the field's first systematic response to this gap. It includes GAN- and restoration-network-generated distortions in its 29,000-image corpus and uses an Elo-style pairwise comparison protocol better suited to ranking subtly different generative outputs than absolute category rating is. That only one of seven major datasets addresses generative artifacts, and that PIPAL itself predates the diffusion-based codecs of Section 3.1-D, is direct empirical evidence for Research Gap 2 (Section 5.1).

4.2. Comparative Analysis

Table 4 consolidates the correlation, reference requirement, generative-artifact sensitivity, and computational-cost characteristics of the metric families reviewed in Sections 2.1 and 2.3. The entries are synthesized directly from the statistical evaluations and ablation studies reported in the primary sources [16], [36], [37], [44], [49], and cross-checked against the architecture-specific evaluation protocols reported for the codecs of Section 3.1.

Table 4. Comparative Characteristics of Representative Quality Metrics

Metric / Codec class	Correlation with human MOS	Reference-free capable	Sensitive to generative hallucination	Computational cost
PSNR	Low–moderate	No	No (insensitive generally)	Very low
SSIM / MS-SSIM	Moderate–high	No	Limited	Low
VIF	High	No	Limited	Moderate
NIQE / BRISQUE	Moderate (distortion-generic)	Yes	Partial (NSS-deviation dependent)	Low–moderate
LPIPS	High	No	Partial (patch-level only)	Moderate (network forward pass)

Metric / Codec class	Correlation with human MOS	Reference-free capable	Sensitive to generative hallucination	Computational cost
DISTS	High	No	Partial (structure/texture split helps)	Moderate
VMAF	High (video, streaming domain)	No	Not evaluated for generative codecs	Moderate
Human MOS / DMOS (BT.500 protocol)	Ground truth by definition	Both FR and NR variants exist	Yes (gold standard)	Very high (panel study)

Three patterns in Table 4 deserve discussion. First, correlation with human MOS increases roughly monotonically from PSNR through structural and information-theoretic metrics to deep-feature metrics. The pattern replicates, at the level of a synthesized cross-study comparison, the individual findings of [16], [49], and matches the historical narrative of Section 1.2. Second, no metric in the table is unambiguously rated "yes" for sensitivity to generative hallucination apart from human MOS itself, even though that failure mode is the one most characteristic of the codec families in Section 3.1-C–D. Section 4.1-C identified this coverage gap, elaborated as Research Gap 2 below. Third, computational cost rises with representational depth (PSNR ≪ SSIM ≈ NIQE ≪ LPIPS/DISTS ≈ VMAF ≪ human panel study), which creates a practical tension between evaluation fidelity and evaluation throughput. The tension becomes acute wherever per-iteration quality feedback is needed during codec training, or where large-scale automated content-pipeline auditing is required, since only the cheapest and least perceptually faithful metrics are tractable at that scale. Section 3.2-B's use-case-conditioned weighting is designed to make this tension visible and manageable, not to resolve it outright.

Fig. 2 (Section 2.1-G) also serves here as a comparative device. Positioning a codec family on the (D, P) operating diagram makes the difficulty visible at a glance. Take a codec optimized for low D, such as the hyperprior and autoregressive family of Section 3.1-A–B, and one optimized for low P, such as the GAN and diffusion family of Section 2.2-C–D. Neither D-type metrics (PSNR, SSIM, MSE) nor P-type metrics (FID, NIQE, human preference) can rank the two fairly on their own. A fair comparison requires reporting both axes, or a use-case-conditioned fusion of both, exactly as proposed in Section 3.2.

5. RESEARCH GAPS, FUTURE DIRECTIONS, AND CONCLUSION

5.1. Research Gaps

Synthesizing the evidence presented in Sections 1.2 to 4.2, this review identifies six concrete, evidence-grounded research gaps that together define the unresolved core of perceptual quality assessment for AI-based image compression.

Gap 1: Absence of a validated semantic-hallucination metric

No metric reviewed in Section 2.3 was designed or validated to detect the class of artifact most characteristic of GAN- and diffusion-based codecs (Section 3.1-C–D), the deep-feature metrics of Section 2.3-D included. The artifact class in question is locally realistic, globally inconsistent synthesized content that alters scene semantics while preserving or even improving patch-level perceptual scores. LPIPS and DISTS operate on local patches and were calibrated on human judgments of classical distortions [16], [17]. NIQE and BRISQUE detect deviations from natural-image statistics that hallucinated content, by construction, may not exhibit [36], [37]. The concern is not hypothetical. HiFiC's own validation required a bespoke human preference study because no existing objective metric was trusted to certify its outputs [19]. The field lacks a metric with the rigor and adoption of SSIM or LPIPS that targets semantic and identity consistency directly.

Gap 2: Subjective datasets lag behind generative codec deployment

Six of the seven major subjective IQA datasets in the reviewed corpus contain only classical, parametrically defined distortions, as Section 4.1-C shows. PIPAL [44], the sole dataset addressing

GAN-generated artifacts, predates the diffusion-based codecs reviewed in Section 3.1-D. No large-scale, publicly available, human-annotated dataset of diffusion-compression artifacts currently exists at a scale and rigor comparable to TID2013 [73] or KADID-10k [54] for classical distortions.

Gap 3: No consensus protocol for reporting rate–distortion–perception operating points

Section 2.1-G and Section 4.2 show that a codec's position on the (D, P) plane is decisive for fair comparison. Yet no standard reporting protocol in the reviewed corpus requires codecs to report both a distortion-type metric and a perception-type metric, still less to report position across a range of rate–distortion–perception operating points reachable through controllable-realism mechanisms [50]. Published rate–distortion curves for different codec families are therefore frequently not comparable on their own terms.

Gap 4: Task fidelity is treated as a separate literature, not an integrated evaluation dimension

Task-oriented compression [69] is an active and empirically well-motivated research direction, as Section 3.1-E shows. Yet no mainstream IQA taxonomy or benchmark integrates task fidelity alongside pixel, structural, and perceptual fidelity as a co-equal evaluation dimension, as Sections 2.3 and 4.1 make clear. Task-oriented compression and perceptual-quality compression remain largely disjoint literatures with separate benchmarks, separate venues, and, most consequentially, separate default metrics, despite both operating on the same underlying rate–distortion–perception surface.

Gap 5: Computational cost of high-fidelity metrics limits their use in training loops and large-scale auditing

Table 4 (Section 4.2) shows a clear inverse relationship between a metric's correlation with human judgment and its computational cheapness. Deep perceptual metrics [16], [17] and human panel studies [32], [33] are the most trustworthy and the least scalable. NR metrics [35]–[37] and PSNR are the most scalable and the least trustworthy for the generative-artifact failure mode of Gap 1. No metric in the reviewed corpus achieves human-competitive correlation and sub-millisecond per-image evaluation cost at the same time, which is what a training-loop reward signal would need at scale.

Gap 6: Standardization efforts have not yet converged on perceptual evaluation criteria for learned codecs

The ongoing JPEG AI standardization effort [63] is the field's first attempt to formally standardize a learned image codec. As of the version of the standard reviewed within this paper's search window, however, no perceptual evaluation protocol analogous to the decades-old ITU-R BT.500 framework [32] has been agreed upon for AI-based codecs. Standards bodies continue to rely on ad hoc combinations of PSNR, SSIM, and informal subjective testing, without a validated multi-dimensional protocol of the kind proposed in Section 3.2.

5.2. Future Research Questions

Motivated directly by the six research gaps of Section 5.1, this review formulates nine future research questions intended to structure the field's next phase of methodological development.

FRQ1: Can a semantic-consistency metric be trained and validated at the same scale and rigor as LPIPS/DISTS, targeting identity- and scene-altering generative hallucination specifically, and how well would it generalize across GAN, diffusion, and future generative decoder families?

FRQ2: What is the minimum scale and diversity of human-annotated data needed to build a PIPAL-scale subjective dataset for diffusion-compression artifacts, and can active-learning or synthetic-augmentation strategies reduce that cost?

FRQ3: Can a standardized rate–distortion–perception reporting protocol, analogous to the Bjøntegaard-delta convention for conventional codec comparison, be defined and adopted community-wide for learned and generative codecs?

FRQ4: How does the optimal fusion weighting of pixel, structural, perceptual, semantic, and task fidelity (Section 3.2-B) vary quantitatively across concrete use cases such as medical imaging, consumer photography, and autonomous perception, and could that weighting itself be learned from use-case-specific outcome data instead of being specified by expert judgment?

FRQ5: Can a single codec be trained to expose a continuously controllable operating point along the full rate–distortion–perception–task surface at inference time, without retraining, extending the universal-representation results of [50] to a four-dimensional setting that includes task fidelity?

FRQ6: What is the computational lower bound for a quality metric that retains human-competitive correlation with MOS while remaining tractable as a per-iteration training signal for large-scale codec optimization?
FRQ7: To what extent do existing no-reference metrics [35]–[37], validated primarily on classical distortions, systematically over- or under-estimate the quality of diffusion-synthesized content, and can that bias be characterized and corrected without full retraining?
FRQ8: What formal evaluation criteria should standards bodies (JPEG AI [63] and its successors) adopt to certify that a learned codec's perceptual claims are reproducible across viewing conditions, display technologies, and cultural and demographic rater populations?
FRQ9: Can the five-dimensional unified framework proposed in Section 3.2 be extended to video and multi-modal (audio-visual) compression, where temporal and cross-modal consistency introduce fidelity dimensions absent from the single-image setting analyzed here?

5.3. Research Roadmap

Table 5 translates the research gaps and questions of Sections 5.1-5.2 into a phased roadmap. It distinguishes three kinds of action: those immediately achievable with existing methods and data, those requiring new data collection or metric development, and those requiring community- and standards-body-level coordination.

Table 5. Short-, Medium-, and Long-Term Research Roadmap

Horizon	Timeframe	Priority actions	Addresses gap(s)
Short-term	0–2 years	(i) Extend PIPAL-style benchmarks to diffusion-generated artifacts; (ii) publish paired (D, P) rate–distortion–perception curves as a reporting norm for new codec papers; (iii) release open-source reference implementations of the unified five-dimension scoring framework (Section 3.2).	Gaps 2, 3
Medium-term	2–5 years	(i) Train and validate a large-scale semantic-consistency / hallucination-detection metric on newly collected generative-artifact data; (ii) develop lightweight distilled versions of deep perceptual metrics suitable for training-loop use; (iii) integrate task-fidelity benchmarks into mainstream IQA leaderboards alongside perceptual benchmarks.	Gaps 1, 4, 5
Long-term	5+ years	(i) Establish a standards-body-endorsed multi-dimensional perceptual evaluation protocol for AI-based codecs, analogous in authority to ITU-R BT.500 for broadcast video; (ii) extend the unified framework to video, multi-modal, and interactive/streaming settings with temporal and cross-modal fidelity dimensions; (iii) develop theoretically grounded, use-case-conditioned fusion weight estimation grounded in decision-theoretic outcome data rather than expert elicitation.	Gaps 3, 4, 6

The roadmap is structured so that short-term actions do not presuppose the outcome of medium- or long-term research. Publishing paired (D, P) curves and adopting the unified scoring framework are achievable with today's metrics and deliver immediate, incremental transparency benefits even before a validated semantic-hallucination metric (medium-term) or a standards-endorsed protocol (long-term) exists. The sequencing is a deliberate methodological choice informed by the historical pattern noted in Section 1.2-D, that evaluation theory in this field has repeatedly lagged engineering practice. The roadmap is designed to narrow that lag rather than reproduce it, by prioritizing

reporting-practice changes that require no new fundamental research alongside the longer-horizon research agenda

5.4. Conclusion

This paper has presented a systematic, critically synthesized review of perceptual quality assessment for AI-based image compression, spanning 1948 to 2023 and grounded in 75 verified sources selected through a PRISMA-informed screening protocol. Three conclusions emerge. First, the trajectory from Shannon's rate–distortion theory through hand-engineered structural metrics to deep-learned perceptual metrics and, most recently, to the formal rate–distortion–perception tradeoff is not a sequence of incremental improvements. It is a series of paradigm shifts in what "quality" is taken to mean. The compression architectures now in widespread deployment, from hyperprior codecs to diffusion-based generative decoders, occupy genuinely different and empirically documented points on the resulting surface. Second, no single existing metric, however sophisticated, is adequate to evaluate this landscape fairly. Pixel-fidelity metrics systematically penalize perceptually preferred generative reconstructions, and even the best deep perceptual metrics were neither designed nor validated to detect the semantic hallucination failure mode specific to generative decoders. Third, and following from the other two, this review has proposed a unified five-dimensional evaluation framework integrating pixel, structural, perceptual, semantic, and task fidelity under a use-case-conditioned adaptive weighting. It is offered as a practical, immediately deployable response, generalizing to AI-based image compression the fused-metric philosophy VMAF pioneered in the video domain, in a setting both broader and more architecturally diverse.

The six research gaps and nine future research questions formulated in Sections 5.1-5.2, together with the phased roadmap of Section 5.3, are offered as a concrete agenda for the research community, for standards bodies, and for practitioners engaged in designing and deploying the next generation of image compression systems. Learned and generative codecs continue to displace conventional standards across consumer, archival, and machine-vision applications. As they do, the adequacy of the metrics used to evaluate them stops being a peripheral methodological concern. It becomes a first-order determinant of whether the field's stated progress, still too often measured by a single number on a single benchmark, reflects real improvement in the properties that matter to the humans and machines who ultimately consume compressed images.

DECLARATIONS:

Acknowledgments	:	Not applicable.
Conflict of Interest	:	The author declares that there is no actual or potential conflict of interest about this article.
Consent to Publish	:	The authors agree to publish the paper in the Global Research Journal of Social Sciences and Management.
Ethical Approval	:	Not applicable.
Funding	:	Author claims no funding was received.
Author Contribution	:	Both the authors confirms their responsibility for the study, conception, design, data collection, and manuscript preparation.
Data Availability Statement	:	The data presented in this study are available upon request from the corresponding author.
Use of Generative Artificial Intelligence	:	AI-assisted tools have been used to improve language editing or grammatical clarity, the authors take full responsibility for the content of the final version.

REFERENCES

[1] G. K. Wallace, "The JPEG still picture compression standard," IEEE Trans. Consum. Electron., vol. 38, no. 1, pp. xviii–xxxiv, Feb. 1992.

[2] D. S. Taubman and M. W. Marcellin, JPEG2000: Image Compression Fundamentals, Standards and Practice. Boston, MA, USA: Kluwer Academic, 2002.

[3] A. Skodras, C. Christopoulos, and T. Ebrahimi, "The JPEG 2000 still image compression standard," IEEE Signal Process. Mag., vol. 18, no. 5, pp. 36–58, Sep. 2001.

[4] Z. Wang and A. C. Bovik, "Mean squared error: Love it or leave it? A new look at signal fidelity measures," IEEE Signal Process. Mag., vol. 26, no. 1, pp. 98–117, Jan. 2009.
[5] J. Ballé, V. Laparra, and E. P. Simoncelli, "End-to-end optimized image compression," in Proc. Int. Conf. Learn. Represent. (ICLR), 2017.
[6] J. Ballé, D. Minnen, S. Singh, S. J. Hwang, and N. Johnston, "Variational image compression with a scale hyperprior," in Proc. Int. Conf. Learn. Represent. (ICLR), 2018.
[7] D. Minnen, J. Ballé, and G. D. Toderici, "Joint autoregressive and hierarchical priors for learned image compression," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 31, 2018, pp. 10771–10780.
[8] D. Minnen and S. Singh, "Channel-wise autoregressive entropy models for learned image compression," in Proc. IEEE Int. Conf. Image Process. (ICIP), 2020, pp. 3339–3343.
[9] Z. Cheng, H. Sun, M. Takeuchi, and J. Katto, "Learned image compression with discretized Gaussian mixture likelihoods and attention modules," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2020, pp. 7939–7948.
[10] [D. He, Y. Zheng, B. Sun, Y. Wang, and H. Qin, "Checkerboard context model for efficient learned image compression," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2021, pp. 14771–14780.
[11] D. He, Z. Yang, W. Peng, R. Ma, H. Qin, and Y. Wang, "ELIC: Efficient learned image compression with unevenly grouped space-channel contextual adaptive coding," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2022, pp. 5718–5727.
[12] B. Bross, Y.-K. Wang, Y. Ye, S. Liu, J. Chen, G. J. Sullivan, and J.-R. Ohm, "Overview of the Versatile Video Coding (VVC) standard and its applications," IEEE Trans. Circuits Syst. Video Technol., vol. 31, no. 10, pp. 3736–3764, Oct. 2021.
[13] [Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," IEEE Trans. Image Process., vol. 13, no. 4, pp. 600–612, Apr. 2004.
[14] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in Proc. 37th Asilomar Conf. Signals, Syst. Comput., vol. 2, 2003, pp. 1398–1402.
[15] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," IEEE Trans. Image Process., vol. 15, no. 2, pp. 430–444, Feb. 2006.
[16] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2018, pp. 586–595.
[17] K. Ding, K. Ma, S. Wang, and E. P. Simoncelli, "Image quality assessment: Unifying structure and texture similarity," IEEE Trans. Pattern Anal. Mach. Intell., vol. 44, no. 5, pp. 2567–2581, May 2022.
[18] E. Agustsson, M. Tschannen, F. Mentzer, R. Timofte, and L. Van Gool, "Generative adversarial networks for extreme learned image compression," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), 2019, pp. 221–231.
[19] F. Mentzer, G. D. Toderici, M. Tschannen, and E. Agustsson, "High-fidelity generative image compression," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 33, 2020, pp. 11913–11924.
[20] L. Theis, T. Salimans, M. D. Hoffman, and F. Mentzer, "Lossy compression with Gaussian diffusion," 2022, arXiv:2206.08889.
[21] E. Hoogeboom, E. Agustsson, F. Mentzer, L. Versari, G. Toderici, and L. Theis, "High-fidelity image compression with score-based generative models," 2023, arXiv:2305.18231.
[22] Y. Blau and T. Michaeli, "The perception-distortion tradeoff," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2018, pp. 6228–6237.
[23] Y. Blau and T. Michaeli, "Rethinking lossy compression: The rate-distortion-perception tradeoff," in Proc. Int. Conf. Mach. Learn. (ICML), vol. 97, 2019, pp. 675–685.
[24] C. E. Shannon, "A mathematical theory of communication," Bell Syst. Tech. J., vol. 27, no. 3, pp. 379–423, 1948.

[25] C. E. Shannon, "Coding theorems for a discrete source with a fidelity criterion," IRE Nat. Conv. Rec., vol. 7, pt. 4, pp. 142–163, 1959.
[26] T. Berger, Rate Distortion Theory: A Mathematical Basis for Data Compression. Englewood Cliffs, NJ, USA: Prentice-Hall, 1971.
[27] T. M. Cover and J. A. Thomas, Elements of Information Theory, 2nd ed. Hoboken, NJ, USA: Wiley, 2006.
[28] V. K. Goyal, "Theoretical foundations of transform coding," IEEE Signal Process. Mag., vol. 18, no. 5, pp. 9–21, Sep. 2001.
[29] J. Alakuijala et al., "JPEG XL next-generation image compression architecture and coding tools," in Proc. SPIE, vol. 11137, Appl. Digit. Image Process. XLII, 2019, Art. no. 111370K.
[30] T. Wiegand, G. J. Sullivan, G. Bjøntegaard, and A. Luthra, "Overview of the H.264/AVC video coding standard," IEEE Trans. Circuits Syst. Video Technol., vol. 13, no. 7, pp. 560–576, Jul. 2003.
[31] G. J. Sullivan, J.-R. Ohm, W.-J. Han, and T. Wiegand, "Overview of the High Efficiency Video Coding (HEVC) standard," IEEE Trans. Circuits Syst. Video Technol., vol. 22, no. 12, pp. 1649–1668, Dec. 2012.
[32] Methodologies for the Subjective Assessment of the Quality of Television Images, Recommendation ITU-R BT.500-14, Int. Telecommun. Union, Geneva, Switzerland, Oct. 2019.
[33] Subjective Video Quality Assessment Methods for Multimedia Applications, Recommendation ITU-T P.910, Int. Telecommun. Union, Geneva, Switzerland, 2021.
[34] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "FSIM: A feature similarity index for image quality assessment," IEEE Trans. Image Process., vol. 20, no. 8, pp. 2378–2386, Aug. 2011.
[35] A. K. Moorthy and A. C. Bovik, "Blind image quality assessment: From natural scene statistics to perceptual quality," IEEE Trans. Image Process., vol. 20, no. 12, pp. 3350–3364, Dec. 2011.
[36] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," IEEE Trans. Image Process., vol. 21, no. 12, pp. 4695–4708, Dec. 2012.
[37] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a 'completely blind' image quality analyzer," IEEE Signal Process. Lett., vol. 20, no. 3, pp. 209–212, Mar. 2013.
[38] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 25, 2012, pp. 1097–1105.
[39] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in Proc. Int. Conf. Learn. Represent. (ICLR), 2015.
[40] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2016, pp. 770–778.
[41] S. Bosse, D. Maniry, K.-R. Müller, T. Wiegand, and W. Samek, "Deep neural networks for no-reference and full-reference image quality assessment," IEEE Trans. Image Process., vol. 27, no. 1, pp. 206–219, Jan. 2018.
[42] H. Talebi and P. Milanfar, "NIMA: Neural image assessment," IEEE Trans. Image Process., vol. 27, no. 8, pp. 3998–4011, Aug. 2018.
[43] E. Prashnani, H. Cai, Y. Mostofi, and P. Sen, "PieAPP: Perceptual image-error assessment through pairwise preference," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2018, pp. 1808–1817.
[44] J. Gu, H. Cai, H. Chen, X. Ye, J. S. Ren, and C. Dong, "PIPAL: A large-scale image quality assessment dataset for perceptual image restoration," in Proc. Eur. Conf. Comput. Vis. (ECCV), 2020, pp. 633–651.
[45] S. Yang, T. Wu, S. Shi, S. Lao, Y. Gong, M. Cao, J. Wang, and Y. Yang, "MANIQA: Multi-dimension attention network for no-reference image quality assessment," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), 2022, pp. 1191–1200.
[46] J. Ballé, V. Laparra, and E. P. Simoncelli, "Density modeling of images using a generalized normalization transformation," in Proc. Int. Conf. Learn. Represent. (ICLR), 2016.

[47] J. Liu, H. Sun, and J. Katto, "Learned image compression with mixed transformer-CNN architectures," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2023, pp. 14388–14397.
[48] S. Santurkar, D. Budden, and N. Shavit, "Generative compression," in Proc. Picture Coding Symp. (PCS), 2018, pp. 258–262.
[49] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," IEEE Trans. Image Process., vol. 15, no. 11, pp. 3440–3451, Nov. 2006.
[50] G. Zhang, J. Qian, J. Chen, and A. Khisti, "Universal rate-distortion-perception representations for lossy compression," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 34, 2021, pp. 11517–11529.
[51] D. Moher, A. Liberati, J. Tetzlaff, D. G. Altman, and The PRISMA Group, "Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement," PLoS Med., vol. 6, no. 7, 2009, Art. no. e1000097.
[52] M. J. Page et al., "The PRISMA 2020 statement: An updated guideline for reporting systematic reviews," BMJ, vol. 372, 2021, Art. no. n71.
[53] B. Kitchenham and S. Charters, "Guidelines for performing systematic literature reviews in software engineering," Keele Univ. and Univ. Durham, EBSE Tech. Rep. EBSE-2007-01, 2007.
[54] H. Lin, V. Hosu, and D. Saupe, "KADID-10k: A large-scale artificially distorted IQA database," in Proc. 11th Int. Conf. Qual. Multimedia Exp. (QoMEX), 2019, pp. 1–3.
[55] G. Toderici, S. M. O'Malley, S. J. Hwang, D. Vincent, D. Minnen, S. Baluja, M. Covell, and R. Sukthankar, "Variable rate image compression with recurrent neural networks," 2015, arXiv:1511.06085.
[56] G. Toderici, D. Vincent, N. Johnston, S. J. Hwang, D. Minnen, J. Shor, and M. Covell, "Full resolution image compression with recurrent neural networks," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2017, pp. 5306–5314.
[57] L. Theis, W. Shi, A. Cunningham, and F. Huszár, "Lossy image compression with compressive autoencoders," in Proc. Int. Conf. Learn. Represent. (ICLR), 2017.
[58] O. Rippel and L. Bourdev, "Real-time adaptive image compression," in Proc. Int. Conf. Mach. Learn. (ICML), vol. 70, 2017, pp. 2922–2930.
[59] M. Li, W. Zuo, S. Gu, D. Zhao, and D. Zhang, "Learning convolutional networks for content-weighted image compression," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2018, pp. 3214–3223.
[60] E. Agustsson, F. Mentzer, M. Tschannen, L. Cavigelli, R. Timofte, L. Benini, and L. Van Gool, "Soft-to-hard vector quantization for end-to-end learning compressible representations," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 30, 2017, pp. 1141–1151.
[61] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in Proc. Int. Conf. Learn. Represent. (ICLR), 2014.
[62] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 30, 2017, pp. 5998–6008.
[63] J. Ascenso, E. Alshina, and T. Ebrahimi, "The JPEG AI standard: Providing efficient human and machine visual data consumption," IEEE MultiMedia, vol. 30, no. 1, pp. 100–111, Jan.–Mar. 2023.
[64] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 27, 2014, pp. 2672–2680.
[65] D. He, Z. Yang, H. Yu, T. Xu, J. Luo, Y. Chen, C. Gao, X. Shi, H. Qin, and Y. Wang, "PO-ELIC: Perception-oriented efficient learned image coding," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), 2022, pp. 1764–1769.
[66] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 33, 2020, pp. 6840–6851.

[67] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2022, pp. 10684–10695.
[68] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 30, 2017, pp. 6626–6637.
[69] L. D. Chamain, F. Racapé, J. Bégaint, A. Pushparaja, and S. Feltman, "End-to-end optimized image compression for machines, a study," in Proc. Data Compress. Conf. (DCC), 2021, pp. 163–172.
[70] Z. Li, A. Aaron, I. Katsavounidis, A. Moorthy, and M. Manohara, "Toward a practical perceptual video quality metric," Netflix Technol. Blog, Jun. 2016.
[71] C. G. Bampis, Z. Li, and A. C. Bovik, "SpatioTemporal feature integration and model fusion for full reference video quality assessment," IEEE Trans. Circuits Syst. Video Technol., vol. 29, no. 8, pp. 2256–2270, Aug. 2019.
[72] N. Ponomarenko, V. Lukin, A. Zelensky, K. Egiazarian, M. Carli, and F. Battisti, "TID2008—A database for evaluation of full-reference visual quality assessment metrics," Adv. Modern Radioelectron., vol. 10, no. 4, pp. 30–45, 2009.
[73] N. Ponomarenko, L. Jin, O. Ieremeiev, V. Lukin, K. Egiazarian, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti, and C.-C. J. Kuo, "Image database TID2013: Peculiarities, results and perspectives," Signal Process., Image Commun., vol. 30, pp. 57–77, Jan. 2015.
[74] V. Hosu, H. Lin, T. Sziranyi, and D. Saupe, "KonIQ-10k: An ecologically valid database for deep learning of blind image quality assessment," IEEE Trans. Image Process., vol. 29, pp. 4041–4056, 2020.
[75] D. Ghadiyaram and A. C. Bovik, "Massive online crowdsourced study of subjective and objective picture quality," IEEE Trans. Image Process., vol. 25, no. 1, pp. 372–387, Jan. 2016.

Authors' Profiles



Mr. Mahesh Boddu obtained his B.Tech in ECE in 2012 and M.Tech in CSE in 2015 both are from JNTU Hyderabad. He started his career in 2012 as Assistant Professor of CSE Department with ABCD, Hyderabad. He worked in various Engineering colleges and continued in teaching up to June 2019. In July 2019 2009 he moved to M/S iQor Services Pvt. Limited as Training Manager. He has worked in Reliance Jio and IFB Appliance heading their respective training Divisions of the state of Andhra Pradesh. He has over 10 years of Industry and 5 Years of Teaching experience. His research interest includes Computer Vision, Image Processing, Machine learning, Artificial intelligence, Deep Learning algorithms, Cloud Computing and IoT. He is also a life member of IETE and ISTE.



Dr. Soumitra Kumar Mandal has obtained his B.E. from Bengal Engineering College (Now IIEST), Shibpur, M.Tech from Institute of Technology, Banaras Hindu University, Varanasi and Ph.D. from Punjab University, Chandigarh all in Electrical Engineering. He started his career as Lecturer at SSGM Engineering College, Shegaon, and then moved to Punjab Engineering College, Chandigarh. In February, 2004, he has been appointed as Assistant Professor of Electrical Engineering in National Institute of Technical Teachers' Training and Research (NITTTR), Kolkata. He is now serving as Professor of Electrical Engineering in the same institute. Throughout his academic career, he has published about 45 research papers in National and International Journals and presented many papers in conferences. He has also published 8 Text books for undergraduate and Post Graduate Students of Electrical Engineering. His research interests include Microprocessor and Microcontroller based System Design, Embedded System Design, Computer Controlled Drives, Neuro-fuzzy Computing, Signal Processing and VLSI design. He is also a life member of ISTE and a member of IE.