INTERNATIONAL JOURNAL OF RESEARCH AND INNOVATION IN APPLIED SCIENCE (IJRIAS)
ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue VII July 2026

Semantic Image Compression for Rate, Distortion, and Performance Optimization: A Systematic Review with Quantitative Synthesis

Manga, I., Mathew, R., & Bali, B

Department of Computer Science, Adamawa State University, Mubi, Nigeria

DOI: https://doi.org/10.51584/IJRIAS.2026.11070021

Received: 10 July 2026; Accepted: 15 July 2026; Published: 28 July 2026

ABSTRACT

The semantic image compression is revolutionizing the intelligent visual communication by compressing the image representation for machine perception and downstream tasks instead of human viewing. Although there have been great improvements with learned compression, semantic communication, and task-aware coding, there are still problems with architecture design, evaluation protocols, benchmark datasets, and deployment. This review, based on the PRISMA framework, examines 54 selected studies published from January 2018 to June 2026. Study-level coding was used to create a common coding scheme to perform a descriptive quantitative evidence synthesis, and frequencies and percentages calculated across learning architectures, optimization objectives, semantic preservation strategies, benchmark datasets, evaluation metrics, and application domains. The results demonstrate the straightforward shift from conventional rate–distortion optimization to semantic and task-aware compression. There are several major difficulties, including lack of standard evaluation frameworks, low cross-domain generalization, low reproducibility, and low unified metrics that combine compression efficiency and semantic fidelity. The review summarizes the existing methodologies, introduces new research trends, and offers useful insights for designing efficient semantic image compression systems for edge intelligence, medical imaging, autonomous platforms and communication networks empowered by artificial intelligence. The emphasis is on what steps can be taken towards learned image compression, which is a subject pursued by the Deep Learning (DL) community.

Keywords: Deep learning, Learned image compression, Rate-distortion optimization, Semantic image compression, Task-aware compression

INTRODUCTION

The field of image compression has shifted from traditional transform coding techniques to data-driven techniques to learn compact image representations from data. The conventional standards, such as JPEG, JPEG 2000 (J2K), HEVC, rely on optimization of rate–distortion (RD) trade-offs for human visual perception. However, the proliferation of Artificial Intelligence (AI), Machine Vision (MV), and Intelligent Communication Systems (ICS) has turned the way of generating, conveying and using visual information upside down. Images mostly used by the human eye, but are increasingly more and more processed on the level of an intelligent system for automatic analysis, decision-making and control.

The last few years have seen remarkable advances in learned image compression technology, such as variational autoencoders and hyperprior-based algorithms that have improved compression efficiency (Minnen et al., 2018). However, novel applications like autonomous systems, remote sensing, medical imaging, intelligent healthcare, edge intelligence and AI-enhanced learning platforms require compressed information representations, which not only accurately reconstruct images but also, more importantly, capture the essential data for subsequent machine learning (ML) tasks. Thus, the semantic-aware image compression is now a hot topic of research that focuses on retaining image semantic information useful for image classification, object detection, segmentation, retrieval and intelligent decision-making support.

The recent growth of both the number of distributed AI systems and the volume of high-resolution visual data has just added to this momentum. Edge computing environments have strict latency, computation and energy constraints (Friha et al., 2024), and bandwidth limitations continue to be a challenge in wireless sensor networks, satellite communications and low-altitude communication infrastructures (Lungisani et al., 2022; Hui et al., 2025). The practical needs have promoted the development of compression methods that provide the highest semantic gain for the lowest communication cost. In recent years, the use of semantic representations has proven to be helpful for RSM, and this has led to the increased relevance of the development of intelligent visual communication (Carlos et al., 2021), explainable AI systems (Bali, 2025), and semantic-aware segmentation (Minaee et al., 2021; Sun et al., 2021; Wu et al., 2024). Similarly, the need for fidelity (Getu et al., 2024; Guo et al., 2024) has been replaced by semantic information needs due to the use of semantic communication.

Nevertheless, there is still a lack of standardization in the field of semantic-aware image compression due to the variety of architectures, optimization goals, datasets, and metrics used. So far, studies that have been conducted dealing with all three objectives are rare. The majority of studies focus on different aspects of compression, and as such, it is difficult to compare and identify best practices. The present reviews mainly focused on specific models, applications or techniques with little regard for the compression efficiency, quality of reconstruction and semantic preservation.

The purpose of this review is to examine recent works on image compression technologies that include semantic information and analyze the comparisons of the various learning architectures, semantic preservation techniques and optimization methods. It helps understand the benefits and drawbacks of the compression quality and precision of the tasks, identify methodological shortcomings and identify trends in the datasets and in the application fields. The review provides a concise overview of techniques, benchmarks, metrics and applications, highlighting challenges such as reproducibility, semantic robustness and multi-objective optimization.

METHODOLOGY

The PRISMA 2020-based systematic review methodology consisted of six phases: identification, screening, eligibility check, quality appraisal, data extraction and qualitative synthesis to review advancements on semantic-aware image compression in AI domain, computer vision, and ICS. The search process is divided into the following steps:

Searching for data sources

The two reviewers independently searched the following databases: Scopus, Web of Science, IEEE Xplore and ACM Digital Library (11 to 20 June 2026) for publications that appeared in the databases between January 2018 and 20 June 2026. Records retrieved were exported in BibTeX format and combined in one reference library. The choice of these databases was due to their large scope of peer-reviewed and conference proceedings research in areas of AI, computer vision, image processing, multimedia systems and communication engineering. All databases were searched under the same search strategy, which included:

("learned image compression" OR "neural image compression" OR "deep image compression" OR "image compression") AND ("semantic compression" OR "semantic communication" OR "task-aware compression" OR "machine vision" OR "perceptual compression")

Automatic deduplication along with manual verification was performed. Studies were screened by two reviewers independently according to a pre-established criteria and in case of disagreement a consensus was reached. Inter-reviewer agreement in the process of title and abstract screening was evaluated through Cohen's κ coefficient with an almost perfect agreement ($\kappa = 0.89$).

Table 1. Summary of the Literature Search Strategy

Database	Search Period	Search String	Final Studies
Scopus	2018–2026	("learned image compression" OR "neural image compression" OR "deep image compression" OR "image compression") AND ("semantic compression" OR "semantic communication" OR	24

INTERNATIONAL JOURNAL OF RESEARCH AND INNOVATION IN APPLIED SCIENCE (IJRIAS)
ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue VII July 2026

		"task-aware compression" OR "machine vision" OR "perceptual compression")	
WoS	2018–2026	Same search strategy	11
IEEE	2018–2026	Same search strategy	15
ACM	2018–2026	Same search strategy	4
Total	2018–2026	—	54

Study Selection

Study selection followed PRISMA 2020 phases: identification, screening, eligibility, and inclusion as shown in Figure 1.

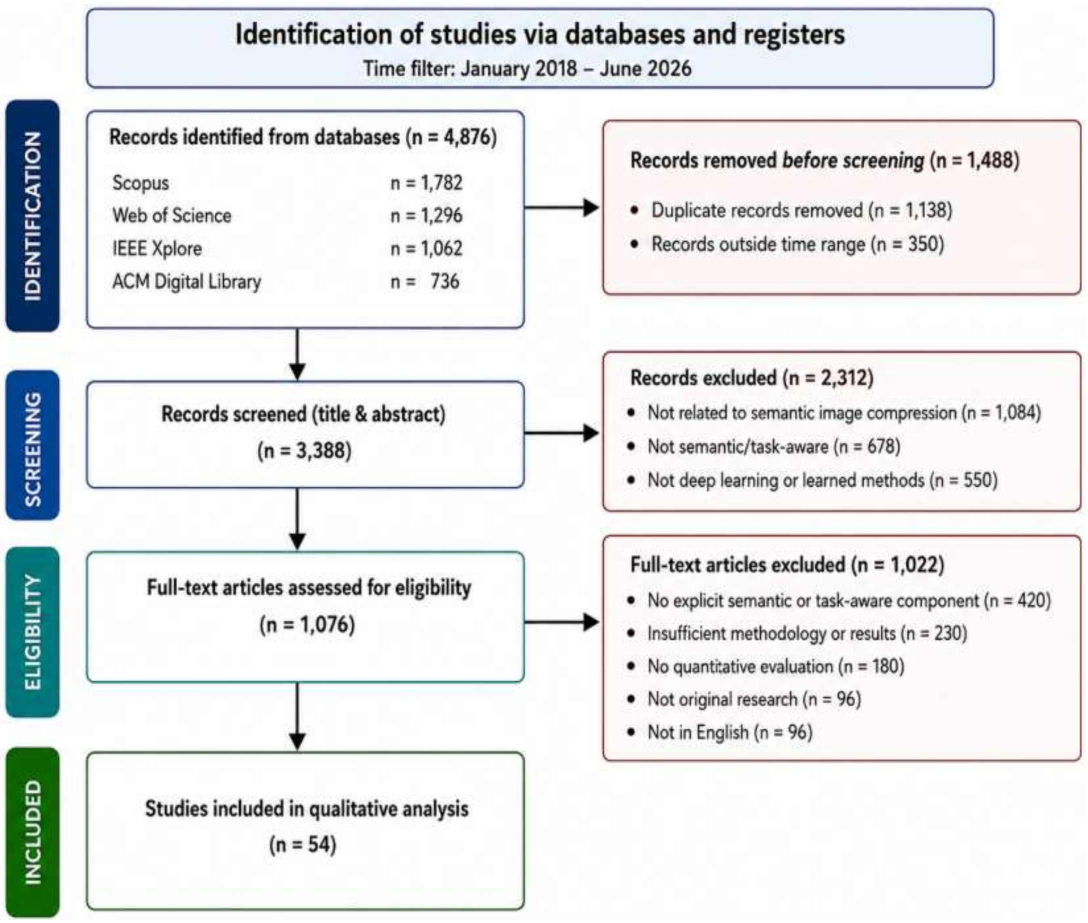

Figure 1: PRISMA flow diagram for the process in selecting studies.

After screening and eligibility assessment, 54 studies were included in the final synthesis.

Eligibility Criteria

The studies were selected, and the inclusion criteria were that they were based on learned/ semantic-aware image compression, included appropriate optimization criteria, included optimization problem or optimization goals that were identifiable, included validated datasets and were published in English in peer-reviewed journals and conference proceedings. Studies that were only about traditional compression, did not have methodological details or quantitative results, did not cite references and were not on the topic of semantic image compression and intelligent visual communication were excluded.

Data Extraction

A data extraction protocol was developed to ensure that all studies extracted data in a standardized manner, considering the variety of studies that could be included. For each publication, the following information was also extracted: Information about the publication; Learning architecture (Convolutional Neural Networks (CNN), variational autoencoders (VAEs), Transformer, GAN or hybrid models); Compression methodology; Benchmark datasets; Evaluation metrics; Compression performance (bitrate or bits-per-pixel); Reconstruction quality metrics Peak Signal-to-Noise Ratio (PSNR), Multi-Scale Structural Similarity Index Measure (MS-SSIM), and SSIM); Downstream task performance (classification accuracy, mean Average Precision (mAP) or Intersection over Union (IoU) scores). Information that was retrieved was used for systematic comparison of methods, data, evaluation practice, application area and reported performance in the literature. Studies were coded using a predefined protocol covering architecture, optimization, datasets, evaluation metrics, application domain, semantic strategy, and limitations. Coding was independently verified by both reviewers before descriptive analysis.

Quality Assessment

Four methodological criteria for evaluating methodological quality were used and each criterion scored on a 0–2 scale (maximum score = 8): (1) dataset quality, (2) experimental rigor, (3) evaluation completeness, and (4) reproducibility. The studies were categorized as low (0–4), moderate (5–6) and high (7–8) quality. Only high and moderate quality studies were included in the final synthesis.

Data Synthesis

Selected studies were synthesized by a systematic comparative approach by considering the architectural structures, semantic strategies, compression goals, datasets, metrics and applications. The analysis of the performance trends, limitations, research gaps and future directions of the semantic-aware image compression was conducted.

Background

The traditional signal processing techniques of image compression have been upgraded to intelligent representation learning by using AI techniques. Traditional applications focused on minimizing storage and transmission expenses, while still preserving a "human quality", while current applications must guarantee that the semantic information is stored for machine interpretation. This change has resulted in the emergence of learned, semantic and task-oriented compression techniques. In classical standards (e.g., JPEG, JPEG 2000, HEVC), transform coding, quantization, prediction and entropy coding are used to reduce the redundancy and, quite importantly, improve the reconstruction quality and compression efficiency. Hence, Mean Squared Error (MSE) and PSNR generally judge the quality of the image based on the distortion, which is evaluated. With the help of better design of transform and adaptive filtering techniques, and optimized quantization techniques (Canterle et al., 2020; Iqbal & Kwon, 2021), the traditional coding methods have been modified. In addition, ML has been integrated into classical coding processes to improve the application of traditional applications in a specific field, such as medical images in medical image compression (Dimililer 2022).

Although widely used, classical compression techniques have some drawbacks in the AI world. This is because they assume that the less distortion at the level of the pixels, the more meaningful content will be retained in the image. This is not the case, however, with machine vision applications, since the success of the application is based on the semantic information and not similarity in the visual content. In previous perceptual image quality studies, many pixel-based quality metrics were found to be inadequate to measure perceptual similarity or semantic relevance (Ahn et al., 2022; Zhang et al., 2018). Classical image compression is accordingly still very suitable for human visual communication, but not for intelligent systems, where information relevant to the task has to be preserved.

DL has revolutionized image compression by introducing end-to-end trainable neural networks to replace the manually designed transforms that have been used before (Bali et al., 2020; Bali, et al., 2026a). For learned image compression, encoder-decoder architectures are used, which learn a compact latent representation for compression and image reconstruction. VAEs and hyperprior models were introduced in Minnen et al. (2018), and it was shown that neural networks can be as competitive as traditional codecs under the same bitrate and can learn efficient entropy models.

Subsequent efforts have greatly enhanced coding efficiency through the use of context-adaptive entropy modelling (Fu et al., 2021), the development of attention mechanisms (Cheng et al., 2020), the creation of hierarchical latent representations (Chen et al., 2025) and the design of lightweight network structures suitable for resource-limited environments (Liu et al., 2024). In addition, some studies address wavelet-inspired neural transforms (Ma et al., 2020), low-complexity transform design (Canterle et al., 2020), recognition-aware restoration (Yang et al., 2023), inpainting-based reconstruction, recurrent quality assessment (Chen et al., 2020), hybrid attention models (Bao et al., 2022), and progressive compression strategies supporting multiple downstream tasks (Lei et al., 2022).

Recently, some machine vision compression learning frameworks have been introduced to integrate the machine vision preprocessing, diffusion-based reconstruction, and hierarchical semantic restoration in large Visual Language Models (LVLs) (Li et al., 2024; Lu et al., 2024; Xu et al., 2025; Li et al., 2025). These developments make it clear that learned image compression is not just about reconstructing images anymore, but about learning to represent them under adaptation, allowing for more advanced applications of AI.

However, the majority of learned compression models are still based on distortion-distortion objective functions for optimizing the reconstruction of the pixels. While these techniques can greatly enhance the compression efficiency, they do not explicitly ensure the preservation of semantic information that is important for follow-on applications. Training instability (Kim et al., 2022), lack of interpretability (Linardatos et al., 2020), and the application of different evaluation protocols make it difficult to make a direct comparison of different approaches. As a result, learned image compression is a great step forward relative to traditional coding, but also carries with it numerous drawbacks of distortion-based optimization.

The idea of semantic compression represents an enhancement of conventional image coding with the aim of transmitting meaningful information than exact reconstruction of the pixels. Unlike the approach of preserving visual content, semantic compression aims to retain an image representation that facilitates inference tasks performed downstream, such as image classification, object detection, semantic segmentation, retrieval and decision support. This paradigm is similar to the semantic communication paradigm, which aims to convey meaning instead of signals (Getu et al., 2024; Guo et al., 2024; Liang et al., 2024). In this context, semantic fidelity is defined as how well compressed representations maintain the information necessary for correct machine interpretation when there are communication constraints.

Multiple solutions have been proposed to help semantic-aware image coding. Saliency-guided methods (Fischer et al., 2023) partition the bits into informative parts of the image, and semantic decomposition (Chang et al., 2023) first separates the information content of the image that is relevant to the task from background information before compressing it.

Semantic segmentation for communication directly incorporates semantic understanding into compression and transmission processes (Wu et al., 2024). This has come on the heels of the rise of foundation models and multimodal learning. Vision-language models learn shared semantic representations across different modalities (Radford et al., 2021); unimodal semantic compression takes advantage of complementary textual and visual information to achieve better transmission efficiency (Shen et al., 2025). In recent years, semantic correction with LLMs has been explored in recent works to enhance robustness under bandwidth constraints, and feature-driven semantic communication has been explored as a way to improve robustness under bandwidth constraints (Guo et al., 2025; Zhang et al., 2025). Although some recent advances have been made, the current situation of heterogeneous architectures, datasets, metrics, and evaluation protocols still poses serious challenges to reproducibility and comparative evaluation.

For lossy image compression, RD theory can be used to control the trade-off between coding efficiency and reconstruction quality, as measured by PSNR, MSE and SSIM. While RD optimization has been used for decades in image compression research, it comes with a number of limitations for AI applications due to the pixel-based fidelity measures it uses. Preserving semantic content does not necessarily mean there is high reconstruction quality, and images with relatively low PSNR can have a great task performance on the downstream side. In addition, deep feature-based perceptual metrics have been shown to confirm that semantic similarity can significantly vary from pixel similarity (Zhang et al., 2018).

Conventional RD optimization has been extended to include perceptual quality, semantic preservation and downstream task performance in compression goals in recent advances in learned image compression (Guerin et al., 2022). But there is still no universal framework to evaluate compression efficiency, visual quality and semantic usefulness in various applications. As a result, it is hard to compare different methods of evaluation because of poorly defined metrics and protocols.

Task-aware compression moves the goal from visual reconstruction towards capturing task-relevant details for efficient downstream tasks and low communication cost. Recent studies have shown that task-aware compression works well for a wide range of AI tasks. Recognition-aware learning enables improved accuracy of image classification in low bit rate conditions and semantic segmentation for compression improves the understanding of the scene during transmission (Wu et al., 2024). Other methods have been proposed in the fields of object detection, remote sensing, multimedia Internet of Things (IoT), mobile vision systems, and intelligent edge computing (Carlos et al., 2021; Noura et al., 2023; Hu et al., 2025). Adaptive Compression Strategies: Emerging foundation models, multimodal architectures, and generative AI further enable adaptive compression strategies that jointly optimize communication efficiency and downstream intelligence (Liang et al., 2024; Liang & Li, 2025; Shen et al., 2025).

Though great strides have been made, there are still obstacles to overcome. Current approaches often use domain-specific data sets, use different evaluation metrics, and have limited generalizability to other domains (Anantrasirichai & Bull, 2022; Liu et al., 2024). Other issues are computational complexity, resource limitations, security, privacy and trustworthiness in semantic communication systems (Won et al., 2024). To overcome these drawbacks, benchmarking, evaluation protocols, and assessment frameworks must be standardized, reproducible and integrated to be able to assess compression efficiency, reconstruction quality and downstream task performance in combination. In general, the development of classical to semantic- and task-aware compression is the transition to intelligent visual representation and forms the basis of the comparative analysis in the subsequent sections.

LITERATURE REVIEW

Semantic image compression involves a variety of learning paradigms that have different architectures, optimization goals and application areas. In this review, the literature is, therefore, organized thematically, allowing for a systematic comparison of the various approaches available. The advent of CNNs laid the groundwork for learned image compression, where the handcrafted transforms were replaced by an end-to-end trainable feature extraction/reconstruction network. Early CNN-based codecs showed that learned representations are able to outperform standard codecs that separately optimize each of the three components: latent feature extraction, entropy modeling, and image reconstruction.

Some representative studies consist of the hyperprior framework proposed by Minnen et al. (2018), attention-guided compression of Cheng et al. (2020), wavelet-inspired neural transforms of Ma et al. (2020), context-based entropy modeling of Fu et al. (2021), and hybrid attention models of Bao et al. (2022). Typical evaluation is done with benchmark datasets like Kodak, DIV2K, COCO and ImageNet data, and the performance is measured in terms of PSNR, MS-SSIM, bits-per-pixel (bpp), and compression ratio.

CNN-based approaches always obtain higher compression efficiency than traditional codecs with a high quality of reconstruction. Their relatively mature architectures, efficient optimization procedures and their ability to leverage existing hardware make them the dominant technique for learned image compression. Although CNN-based models have achieved good performance, they focus mainly on the reconstruction accuracy instead of semantic preservation. Limited modeling in context and reliance on large datasets annotated with human efforts restrict the downstream machine vision performance of their system.

The hyperprior models are a significant improvement in learned image compression, where adaptive latent probability distributions lead to improved entropy coding. Minnen et al. (2018) used a hyperprior design to improve over traditional codecs; this architecture has since become the paradigm for variational hyperpriors. Minnen et al. (2018), who added joint autoregressive and hierarchical before enhancing context modeling, further extended this work. Later, context-adaptive entropy models (Fu et al., 2021), rate-constrained optimization algorithms (Guerin et al., 2022), successive learned compression instability analysis (Kim et al., 2022), and adaptive parameter estimation methods (Chen et al., 2025) were added.

In most hyperprior methods, performance is assessed on Kodak, Tecnick, DIV2K and CLIC datasets, and the following are the main evaluation metrics: PSNR, MS-SSIM, bpp and RD curves. Hyperprior models offer benefits in terms of entropy coding and rate–distortion performance, but most importantly, in terms of semantic preservation. They are also limited in their use in resource-limited environments due to their computational complexity.

Self-attention mechanisms in transformer architectures boost learned image compression by allowing for global context modeling, thus improving semantic representation. Recent works making use of Transformers with applications to image restoration include hierarchical semantic restoration (Li et al., 2025), machine vision compression (Lu et al., 2024), large visual language models (Li et al., 2024), diffusion-assisted perceptual compression (Xu et al., 2025), enhanced structural fidelity (Zhou et al., 2024), and semantic-aware restoration for multiple degradations (Yang et al., 2023). These methods are usually tested on the ImageNet, COCO, Kodak, DIV2K and application-specific datasets, with the PSNR, MS-SSIM, LPIPS, FID, classification accuracy and mAP metrics used to compare them. In general, transformer models outperform CNN-based architectures in holding on to the context and enabling downstream AI tasks. In particular, they are excellent for long-range semantic relations, which are useful for intelligent visual communication and multimodal reasoning.

Although they offer benefits, Transformer-based compression models are demanding in terms of compute and use varying evaluation protocols, which hinder edge deployment and comparative studies. Semantic-aware compression redefines the optimization goal from pixels to information that is used for downstream inferences. These techniques do not aim at just maximizing the visual similarity, but rather explicitly represent semantic features that are relevant for ML applications. Some representative approaches are semantic-aware visual decomposition (Chang et al., 2023), recognition-aware learned compression, semantic segmentation-based communication (Wu et al., 2024), joint perception-distortion enhancement (Duan et al., 2022), progressive semantic compression (Lei et al., 2022), and task-driven semantic-aware image compression (Zhou et al., 2024).

Semantic-aware compression is usually assessed in the COCO, ImageNet, Cityscapes and remote sensing datasets by means of compression (PSNR, MS-SSIM, bpp) and task-level metrics (accuracy, mAP, IoU). Although these approaches can improve the performance of a downstream system, there are still significant issues with lack of common datasets, optimization routines, and assessment protocols that make the system difficult to reproduce and standardize.

Recent advances in foundation models have enabled a new generation of semantic image compression systems that learn rich multimodal representations from large-scale datasets. Beyond conventional image reconstruction, vision-language models, large language models, and generative AI provide semantic priors that improve representation quality and downstream task performance. Representative studies include CLIP-based representation learning (Radford et al., 2021), multimodal semantic communication (Liang et al., 2024), semantic communication with large language models (Guo et al., 2025), image generation for semantic communication (Liang & Li, 2025), and multimodal foundation-model compression (Shen et al., 2025).

By integrating visual, textual, and semantic information, multimodal compression enables context-aware reconstruction and enhanced semantic preservation beyond unimodal approaches. Related advances include point-cloud semantic communication (Huang et al., 2023) and feature-driven semantic communication (Zhang et al., 2025). Despite these improvements, challenges remain in computational efficiency, explainability, security, privacy, reproducibility, and the lack of standardized benchmarks. Overall, the field has evolved from

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue VII July 2026

distortion-based coding toward semantic- and task-driven compression, although inconsistent datasets and evaluation protocols continue to limit objective comparison across studies.

Comparative Analysis

This comparative analysis compares different semantic-aware image compression techniques, their architectures, semantic strategies, benchmark datasets, evaluation metrics, advantages, drawbacks, and prospects for future research. Based on the amount of semantic representation used in the compression process, there are five types of existing methods. Perceptual compression is used to obtain reconstruction quality with perceptual losses without any semantics modelling. Representative methods use deep perceptive similarity measures and attention-guided reconstruction to increase perceptual fidelity (Zhang et al., 2018; Ahn et al., 2022; Bao et al., 2022). They are a good quality to look at, but they are not good for downstream AI tasks.

Feature-preserving compression: stores intermediate features rather than pixel level features. To retain discriminative representations for machine vision (Minnen et al., 2018; Cheng et al., 2020; Wang et al., 2022; Noura et al., 2023). But these techniques are generally based on the pre-trained feature extractors. Task-aware compression optimizes image compression and downstream tasks (including image classification, object detection and semantic segmentation) together. Representative methods, such as recognition-aware compression and progressive learned compression (Lei et al., 2022; Zhou et al., 2024; Zhang et al., 2025; Tang et al., 2026), enhance downstream performance, but may lack generalization to other tasks.

Semantic coding is the most abstracted representation because it represents the semantic units instead of the pixels. Some representative studies are semantic decomposition and semantic communication frameworks (Duan et al., 2022, Chang et al., 2023, Getu et al., 2024, Guo et al., 2024, Huang et al., 2023). These methods are very effective for specific communication tasks, but are application dependent and are costly in computations. Attention-based semantic compression assigns the coding resources adaptively based on the importance of semantics. The representative approaches include attention-guided learned compression, saliency-aware coding, perceptually important neural compression (Cheng et al., 2020; Qin et al., 2022; Bao et al., 2022; Fischer et al., 2023; Li et al., 2025). These methods are effective at efficiently coding the semantically relevant areas while retaining extra computational complexity.

In summary, this taxonomy highlights the progression of image compression techniques from pixel-level coding to semantic representation learning, showing how the field of image communication has become increasingly intelligent in recent years, particularly in the context of AI applications. This trend is also reflected in recent studies. In semantic segmentation, for instance, Wu et al. (2024) proposed a semantic communication framework for semantic segmentation on Cityscapes that boosts segmentation accuracy; Tang et al. (2025), however, proposed a multitask learned compression framework for ImageNet that achieves better performance on downstream tasks while maintaining comparable compression performance.

Table 2. Comparative Analysis of Representative Semantic-Aware Image Compression Methods

Author	Model	Dataset	Compression Objective	Evaluation Metrics	Advantages	Limitations
Minnen et al. (2018)	Joint autoregressive prior	Kodak	Entropy modelling	PSNR, MS-SSIM	Improved compression efficiency	High decoding complexity
Cheng et al. (2020)	Attention-based CNN	Kodak, CLIC	Attention-guided compression	PSNR, MS-SSIM	Better perceptual quality	Increased computation
Lei et al. (2022)	Progressive compression	ImageNet	Task-aware optimization	Accuracy, PSNR	Supports classification	Limited transferability

					tasks	
Chang et al. (2023)	Semantic decomposition	COCO	Semantic coding	mAP, PSNR	Preserves semantic regions	Computationally intensive
Fischer et al. (2023)	Saliency-driven coding	COCO	Machine-oriented coding	Accuracy, bpp	Efficient semantic allocation	Limited benchmark evaluation
Wu et al. (2024)	Semantic segmentation compression	Cityscapes	Semantic communication	IoU, PSNR	High segmentation accuracy	Domain specific
Li et al. (2024)	Vision-language compression	ImageNet	Large-model compression	PSNR, LPIPS	Supports foundation models	High computational cost
Shen et al. (2025)	Multimodal semantic compression	Multi-domain	Foundation model compression	LPIPS, Accuracy	Excellent semantic preservation	Early-stage research
Tang et al. (2025)	Multi-task learned compression	ImageNet	Multi-task semantic optimization	Accuracy, mAP	Strong downstream performance	Limited standardization

From the comparison, it can be seen that the semantic-oriented optimization is replacing the distortion-driven image coding. Previous research has mainly focused on the quality of reconstruction, while the performance of ML has become part of the process of evaluating models in recent research. Evaluation of performance has consequently evolved from pixel-level fidelity to multidimensional assessment. PSNR remains the most widely reported metric for reconstruction accuracy, with hyperprior and Transformer-based models achieving superior rate-distortion performance at comparable bitrates (Minnen et al., 2018; Cheng et al., 2020). For perceptual quality, MS-SSIM better reflects human visual perception and consistently demonstrates the superiority of learned compression over conventional codecs (Cheng et al., 2020; Chen et al., 2025).

Equally, LPIPS has become a standard metric for perceptual similarity by measuring deep feature representations rather than pixel differences, with diffusion-based reconstruction and foundation models achieving improved semantic preservation (Xu et al., 2025; Shen et al., 2025). For machine vision applications, task-specific metrics such as classification accuracy, mAP, and IoU have become essential, as recognition-aware and semantic-aware compression consistently outperform distortion-based methods by explicitly preserving task-relevant information (Chang et al., 2023; Wu et al., 2024; Tang et al., 2025). Nevertheless, inconsistent evaluation protocols and benchmark datasets continue to limit fair comparison and standardized assessment.

Four main research trends are shaping semantic-aware image compression. First, Transformer architectures are increasingly replacing conventional CNNs because of their ability to capture long-range contextual dependencies and semantic relationships (Li et al., 2025; Xu et al., 2025). Second, foundation models and generative AI are advancing semantic communication through multimodal representation learning and context-aware reconstruction (Radford et al., 2021; Liang et al., 2024; Guo et al., 2025; Shen et al., 2025; Yao et al., 2026). Third, growing interest in Edge AI has accelerated research on lightweight compression methods that satisfy bandwidth, latency, energy, and inference constraints for resource-constrained devices (Friha et al., 2024; Liu et al., 2024; Hu et al., 2025). Finally, reproducibility and open science have received increased attention through the adoption of public benchmark datasets, including Kodak, DIV2K, ImageNet, COCO, CLIC, and Cityscapes, and open-source implementations. However, fragmented codebases, evaluation

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue VII July 2026

methodologies, and benchmark protocols continue to hinder independent verification and objective comparison.

DISCUSSION

In image compression area, the compression based on pixels has evolved into smart visual communication that retains information for subsequent MV applications. Yet, there are still methodological and practical issues to be addressed. This section addresses the strengths, weaknesses and future research directions that arose from the analyzed literature.

A major success of recent work is the successful application of DL to image compression. The superior RD performance of VAEs, hyperprior models, attention mechanisms and Transformer architectures over conventional codecs like JPEG and JPEG2000 has been shown consistently. The hyperprior-based models presented by Minnen et al. (2018) set a new standard for learned image compression by jointly optimizing latent representations and entropy models. The reconstruction quality and bitrate efficiency were further improved by subsequent improvements with attention mechanisms (Cheng et al., 2020).

Now, the second major challenge is changed from distortion- based optimization to task-based learning. Recognition-aware compression, semantic decomposition (Chang et al., 2023), and task-driven compression frameworks (Zhou et al., 2024; Tang et al., 2025) highlight the potential of preserving semantic attributes to support a variety of downstream applications, including image classification, object detection, and semantic segmentation tasks with competitive compression capabilities.

Recent researches shows the increasing significance of semantic communication. Instead of transmitting all visual data, semantic communications are tailored to preserve intelligence in information that is beneficial for guiding decision-making, which results in the removal of information from communication without causing severe degradation to the task performance (Getu et al., 2024; Guo et al., 2024). This strategy finds good use in bandwidth-limited application scenarios, including wireless sensor networks, edge intelligence, and satellite communications.

Transformer architectures represent another major milestone by capturing long-range contextual dependencies through self-attention mechanisms, leading to improved compression performance in complex scenes containing multiple semantic objects (Li et al., 2025; Xu et al., 2025). More recently, semantic compression has expanded beyond image reconstruction toward context-aware visual understanding through the integration of vision-language representations, multimodal learning, and foundation models (Radford et al., 2021; Liang et al., 2024; Shen et al., 2025).

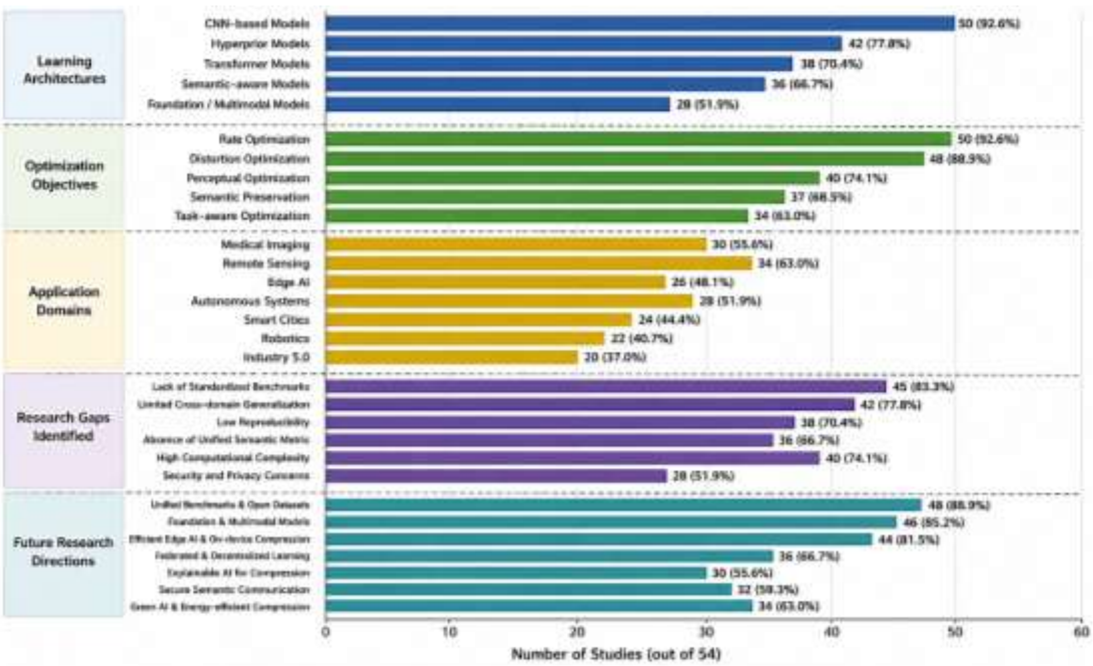

Figure 2. Evidence synthesis of semantic image compression research (2018–2026)

Figure 2 presents an evidence synthesis covering learning architectures, optimization objectives, benchmark datasets, application domains, evaluation metrics, research gaps, and future research directions based on the 54 included studies. Frequencies refers to the number of studies reporting each characteristic and totals exceed 100% because individual studies contributed to multiple categories. Overall, the synthesis is interpreted as an ongoing transition from conventional rate-distortion optimization toward semantic- and task-aware image compression.

While all these developments have come, a number of obstacles still exist. Current research is mainlybased on benchmark datasets such as COCO, DIV2K, Kodak, and ImageNet, limiting cross-domain generalization. In addition, inconsistent datasets, compression frameworks, training strategies, and evaluation metrics hinder fair comparison and reproducibility. Existing models also face high computational complexity and difficulties in deployment on resource-constrained edge devices. Achieving an effective balance between bitrate, reconstruction quality, and downstream task performance while ensuring robustness against adversarial attacks and protecting semantic privacy remains an important research challenge (Won et al., 2024; Guo et al., 2024).

Future research should emphasize foundation models, multimodal semantic communication, adaptive agentic compression, lightweight Edge AI, and hardware-aware optimization. Vision-language models and generative AI have the potential to provide richer semantic representations with lower communication overhead (Radford et al., 2021; Liang et al., 2024; Shen et al., 2025; Li et al., 2026), while lightweight architectures will improve deployment on edge devices (Friha et al., 2024; Liu et al., 2024). In addition, federated learning, Green AI, Explainable AI, and secure semantic communication are expected to improve privacy, interpretability, energy efficiency, and trustworthy deployment in future intelligent communication systems (Linardatos et al., 2020; Won et al., 2024; Guo et al., 2024).

Semantic-aware image compression has significant application potential across multiple domains. In healthcare, it can reduce communication overhead in telemedicine, AI-assisted diagnosis, and cloud-based clinical care while preserving diagnostically relevant information (Dimililer, 2022; Bali, 2025). Interpretable semantic representations supported by Explainable AI will further strengthen clinical trust and decision-making (Linardatos et al., 2020). Beyond healthcare, semantic-aware compression improves communication efficiency for autonomous systems, remote sensing, satellite communications, environmental monitoring, disaster management, smart cities, robotics, and Industry 5.0 by preserving task-relevant information while reducing transmission latency and bandwidth requirements (Lei et al., 2022; Wu et al., 2024; Carlos et al., 2021; Getu et al., 2024; Hui et al., 2025; Bali et al., 2026b; Friha et al., 2024; Guo et al., 2024; Liu et al., 2024). Collectively, these applications demonstrate that semantic-aware image compression is becoming a foundational technology for next-generation intelligent visual communication systems.

CONCLUSION

This review synthesized research on semantic-aware image compression published from 2018 to June 2026 using a PRISMA-guided approach. The findings demonstrate a clear transition from conventional reconstruction-oriented image compression toward semantic and task-aware frameworks that support downstream machine vision applications. Recent advances in DL, including hyperprior models, Transformer architectures, foundation models, and multimodal semantic communication, have substantially improved compression efficiency while preserving information that is essential for intelligent decision-making. Despite these advances, important challenges remain, including dependence on benchmark datasets, limited cross-domain generalization, inconsistent evaluation methods, lack of standardized benchmarks, reproducibility issues, and the computational demands of advanced models. Addressing these challenges will be essential for enabling robust real-world utilization. Emerging technologies such as foundation models, lightweight Edge AI, federated learning, Explainable AI, Green AI, and secure semantic communication provide promising directions for developing more efficient, interpretable, scalable, and trustworthy compression systems. By consolidating current advances, identifying research gaps, and outlining future research priorities, this review provides a comprehensive reference for researchers and practitioners working toward the next generation of intelligent visual communication systems.

REFERENCES

1. Ahn, N., Kang, B., & Sohn, K. A. (2022). Efficient deep neural network for photo-realistic image super-resolution. Pattern Recognition, 127, 108649. https://doi.org/10.1016/j.patcog.2022.108649
2. Anantrasirichai, N., & Bull, D. (2022). Artificial intelligence in the creative industries: A review. Artificial Intelligence Review, 55(1), 589–656. https://doi.org/10.1007/s10462-021-10039-7
3. Bali, B. (2025). Cognitively-informed explainable AI for education: Enhancing student engagement and transparent assessment through human-centered modeling. Journal of Social and Scientific Education, 2(3), 173–190. https://doi.org/10.58230/josse.v2i3.351
4. Bali, B., Nathan, N., & Nzadon, D. T. (2020). A study on medical applications of artificial neural networks. International Journal of Research and Analytical Reviews (IJRAR), 7(3).
5. Bali, B., Ibrahim, M., & Kwaha, F. I. (2026a). Intelligent medical imaging systems for rapid real-time diagnosis: Deep learning, edge AI, hardware acceleration, and healthcare intelligence**.** Advances in Medical Sciences and Research, 1(2), 101–119. https://doi.org/10.55578/amsr.2605.008
6. Bali, B., Goni, I., Manga, I., Kamureyina, E., Ezra, B., & Kwaha, F. I. (2026). Artificial intelligence driven aeromagnetic and satellite data fusion for mineral and groundwater resource mapping in sub-Saharan Africa**.** Discover Geoscience, 4, Article 220. https://doi.org/10.1007/s44288-026-00593-4
7. Bao, Y., Tao, Y., & Qian, P. (2022). Image compression based on hybrid domain attention and postprocessing enhancement. Computational Intelligence and Neuroscience, 2022(1), 4926124. https://doi.org/10.1155/2022/4926124
8. Canterle, D. R., da Silveira, T. L., Bayer, F. M., & Cintra, R. J. (2020). A multiparametric class of low-complexity transforms for image and video coding. Signal Processing, 176, 107685.
9. Carlos, H., Aranda, R., Rivera-De Velasco, M., Rodriguez-Gonzalez, A. Y., & Méndez-López, M. E. (2021). Fishing gear pattern recognition by including supervised autoencoder dimensional reduction. IEEE Geoscience and Remote Sensing Letters, 19, 1-5.
10. Chang, J., Zhang, J., Li, J., Wang, S., Mao, Q., Jia, C., ... & Gao, W. (2023). Semantic-aware visual decomposition for image coding. International Journal of Computer Vision, 131(9), 2333-2355.
11. Chen, L. H., Bampis, C. G., Li, Z., Krasula, L., & Bovik, A. C. (2025). Estimating the resize parameter in end-to-end learned image compression. Signal Processing: Image Communication, 135, 117277.
12. Chen, P., Li, L., Ma, L., Wu, J., & Shi, G. (2020, October). RIRNet: Recurrent-in-recurrent network for video quality assessment. In Proceedings of the 28th ACM international conference on multimedia (834-842).
13. Cheng, Z., Sun, H., Takeuchi, M., & Katto, J. (2020). Learned image compression with discretized Gaussian mixture likelihoods and attention modules. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (7939–7948). https://doi.org/10.48550/arXiv.2001.01568
14. Dimililer, K. (2022). DCT-based medical image compression using machine learning. Signal, Image and Video Processing, 16(1), 55–62. https://doi.org/10.1007/s11760-021-01951-0
15. Duan, S., Chen, H., & Gu, J. (2022). JPD-SE: High-level semantics for joint perception-distortion enhancement in image compression. IEEE Transactions on Image Processing, 31, 4405-4416.
16. Friha, O., Ferrag, M. A., Kantarci, B., Cakmak, B., Ozgun, A., & Ghoualmi-Zine, N. (2024). LLM-based edge intelligence: A comprehensive survey on architectures, applications, security and trustworthiness. IEEE Open Journal of the Communications Society. https://doi.org/10.1109/OJCOMS.2024.3456549
17. Fischer, K., Brand, F., Blum, C., & Kaup, A. (2023, June). Saliency-driven hierarchical learned image coding for machines. In ICASSP 2023 – IEEE International Conference on Acoustics, Speech and Signal Processing (pp. 1–5). IEEE.
18. Fu, H., Liang, F., Lei, B., Zhang, Q., Liang, J., Tu, C., & Zhang, G. (2021). An extended context-based entropy hybrid modeling for image compression. Signal Processing: Image Communication, 95, 116244.
19. Getu, T. M., Kaddoum, G., & Bennis, M. (2024). A survey on goal-oriented semantic communication: Techniques, challenges, and future directions. IEEE Access. https://doi.org/10.1109/ACCESS.2024.3381967

20. Guerin, N. D., da Silva, R. C., & de Oliveira, M. C. (2022). Rate-constrained learning-based image compression. Signal Processing: Image Communication, 101, Article 116544. https://doi.org/10.1016/j.image.2022.116544
21. Guo, S., Wang, Y., Zhang, N., Su, Z., Luan, T. H., Tian, Z., & Shen, X. (2024). A survey on semantic communication networks: Architecture, security, and privacy. IEEE communications surveys & tutorials, 27(5), 2860-2894.
22. Guo, S., Wang, Y., Ye, J., Zhang, A., Zhang, P., & Xu, K. (2025). Semantic importance-aware communications with semantic correction using large language models. IEEE Transactions on Machine Learning in Communications and Networking, 3, 232-245.
23. Hu, B., You, J., Huang, K., Xu, M., Liu, D., & Ma, S. (2025). EffTEE: Efficient image classification and object detection on mobile devices using trusted execution environments. IEEE Access. https://doi.org/10.1109/ACCESS.2025.3538712
24. Huang, Y., Bai, B., Zhu, Y., Qiao, X., Su, X., Yang, L., & Zhang, P. (2023). ISCom: Interest-aware semantic communication scheme for point cloud video streaming on metaverse XR devices. IEEE Journal on Selected Areas in Communications, 42(4), 1003–1021.
25. Hui, M., Zhai, S., Wang, D., Hui, T., Wang, W., Du, P., & Gong, F. (2025). A review of leo satellite communication payloads for integrated communication, navigation, and remote sensing: Opportunities, challenges, future directions. IEEE Internet of Things Journal. https://doi.org/10.1109/JIOT.2025.3553942
26. Iqbal, Y., & Kwon, O. J. (2021). Improved JPEG coding by filtering 8×8 DCT blocks. Journal of Imaging, 7(7), 117. https://doi.org/10.3390/jimaging7070117
27. Kim, J. H., Jang, S., Choi, J. H., & Lee, J. S. (2022). Successive learned image compression: Comprehensive analysis of instability. Neurocomputing, 506, 12-24.
28. Lei, Z., Duan, P., Hong, X., Mota, J. F., Shi, J., & Wang, C. X. (2022). Progressive deep image compression for hybrid contexts of image classification and reconstruction. IEEE Journal on Selected Areas in Communications, 41(1), 72-89.
29. Liang, C., Du, H., Sun, Y., Niyato, D., Kang, J., Zhao, D., & Imran, M. A. (2024). Generative AI-driven semantic communication networks: Architecture, technologies, and applications. IEEE Transactions on Cognitive Communications and Networking, 11(1), 27-47.
30. Liang, C., & Li, D. (2025). Image generation with supervised selection based on multimodal features for semantic communications. IEEE Transactions on Communications.
31. Li, B., Wang, S., Wang, S., & Ye, Y. (2024). High efficiency image compression for large visual-language models. IEEE Transactions on Circuits and Systems for Video Technology. https://doi.org/10.1109/TCSVT.2024.3488181
32. Li, B., Yang, X., Duan, S., & Wang, N. (2026). Toward universal semantic communication via matchable semantic subspace transmission. IEEE Transactions on Image Processing, 35, 5003–5016. https://doi.org/10.1109/TIP.2026.3690331
33. Li, S., Zhang, Z., Xu, M., Jiang, L., Liu, Y., & Zhu, C. (2025). Hierarchical Semantic Compression for Consistent Image Semantic Restoration. IEEE Transactions on Image Processing.
34. Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2020). Explainable AI: A review of machine learning interpretability methods. Entropy, 23(1), 18. https://doi.org/10.3390/e23010018
35. Liu, H. I., Galindo, M., Xie, H., Wong, L. K., Shuai, H. H., Li, Y. H., & Cheng, W. H. (2024). Lightweight deep learning for resource-constrained environments: A survey. ACM Computing Surveys, 56(10), Article 1–42. https://doi.org/10.1145/3657282
36. Lu, G., Ge, X., Zhong, T., Hu, Q., & Geng, J. (2024). Preprocessing enhanced image compression for machine vision. IEEE Transactions on Circuits and Systems for Video Technology. https://doi.org/10.1109/TCSVT.2024.3441049
37. Lungisani, B. A., Lebekwe, C. K., Zungeru, A. M., & Yahya, A. (2022). Image compression techniques in wireless sensor networks: A survey and comparison. IEEE Access, 10, 82,511–82,530. https://doi.org/10.1109/ACCESS.2022.3195891
38. Ma, H., Liu, D., Yan, N., Li, H., & Wu, F. (2020). End-to-end optimized versatile image compression with wavelet-like transform. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(3), 1247–1263. https://doi.org/10.1109/TPAMI.2020.2981717

39. Minaee, S., Boykov, Y., Porikli, F., Plaza, A., Kehtarnavaz, N., & Terzopoulos, D. (2021). Image segmentation using deep learning: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(7), 3523–3542. https://doi.org/10.1109/TPAMI.2021.3059968
40. Minnen, D., Ballé, J., & Toderici, G. D. (2018). Joint autoregressive and hierarchical priors for learned image compression. In Advances in Neural Information Processing Systems, 31.
41. Noura, H. N., Azar, J., Salman, O., Couturier, R., & Mazouzi, K. (2023). A deep learning scheme for efficient multimedia IoT data compression. Ad Hoc Networks, 138, 102998. https://doi.org/10.1016/j.adhoc.2022.102998
42. Qin, H., Han, L., Wang, J., Zhang, C., Li, Y., Li, B., & Hu, W. (2022, October). Attention-aware learning for hyperparameter prediction in image processing pipelines. In European Conference on Computer Vision (271–287). Cham, Switzerland: Springer Nature.
43. Radford, A., Kim, J W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021, July). Learning transferable visual models from natural language supervision. In International Conference on Machine Learning (8748–8763). PMLR.
44. Shen, R., Wu, H., Zhang, W., Hu, J., & Gunduz, D. (2025, August). Compression beyond pixels: Semantic compression with multimodal foundation models. In 2025 IEEE 35th International Workshop on Machine Learning for Signal Processing (MLSP) (1-6). IEEE.
45. Sun, T., Long, L., Yang, R., Li, K., & Liang, J. (2021). Extended virtual signal injection control for MTPA operation of IPMSM drives with online derivative term estimation. IEEE Transactions on Power Electronics, 36(9), 10602-10611.
46. Tang, R., Zhou, P., Wang, K., Yan, J., Shi, C., & Li, Y. (2026). SeGD: A plug-and-play semantic discriminator for task-oriented image compression. Neurocomputing, 685, Article 133667. https://doi.org/10.1016/j.neucom.2026.133667
47. Tang, Z., Yi, X., & Wang, H. (2025). Toward Learned Image Compression for Multiple Semantic Analysis Tasks. IEEE Transactions on Broadcasting.
48. Wang, C., Shen, Q., Wang, X., & Jiang, G. (2022). Momentum feature comparison network based on generative adversarial network for single image super-resolution. Signal Processing: Image Communication, 106, 116726.
49. Won, D., Woraphonbenjakul, G., Wondmagegn, A. B., Tran, A. T., Lee, D., Lakew, D. S., & Cho, S. (2024). Resource management, security, and privacy issues in semantic communications: A survey. IEEE Communications Surveys & Tutorials, 27(3), 1758-1797.
50. Wu, J., Wu, C., Lin, Y., Yoshinaga, T., Zhong, L., Chen, X., & Ji, Y. (2024). Semantic segmentation-based semantic communication system for image transmission. Digital Communications and Networks, 10(3), 519-527.
51. Xu, T., Li, J., Li, B., Wang, Y., Zhang, Y. Q., & Lu, Y. (2025). Picd: Versatile perceptual image compression with diffusion rendering. In Proceedings of the Computer Vision and Pattern Recognition Conference (pp. 28436-28445).
52. Yang, Z., Huang, J., Chang, J., Zhou, M., Yu, H., Zhang, J., & Zhao, F. (2023). Visual recognition-driven image restoration for multiple degradation with intrinsic semantics recovery. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (14059-14070).
53. Yao, L., Wei, Y., Zhou, Y., Li, Z., Chi, N., & Wong, K.-K. (2026). You only transmit once: Unified generation and comprehension for efficient semantic communication. IEEE Journal on Selected Areas in Communications. Advance online publication. https://doi.org/10.1109/JSAC.2026.3655643
54. Zhang, H., Fan, S., Zou, S., Yu, Z., & Zheng, B. (2024). Deep underwater image compression for enhanced machine vision applications. Frontiers in Marine Science, 11, 1411527.
55. Zhang, R., Isola, P., Efros, A. A., Shechtman, E., & Wang, O. (2018). The unreasonable effectiveness of deep features as a perceptual metric. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (586–595). https://doi.org/10.1109/CVPR.2018.00068
56. Zhang, J., Zhang, Y., Ji, B., Chen, A., Liu, A., & Xu, H. (2025). Feature-driven semantic communication for efficient image transmission. Entropy, 27(4), 369.
57. Zhou, F., Huang, X., Zhang, P., Wang, M., Wang, Z., Zhou, Y., & Yin, H. (2024, October). Enhanced screen content image compression: A synergistic approach for structural fidelity and text integrity preservation. In Proceedings of the 32nd ACM International Conference on Multimedia (7900-7908).

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue VII July 2026

58. Zhou, J., Xiao, H., Li, D., Wang, L., Liang, P., Zhou, J., & Yang, Y. (2024, November). Task-Driven Semantic-Aware Image Compression. In International Conference on Ubiquitous Computing and Ambient Intelligence (442-453). Cham: Springer Nature Switzerland.