frontiers | Frontiers in Computer Science

TYPE Mini Review
PUBLISHED 18 September 2025
DOI 10.3389/fcomp.2025.1626641

OPEN ACCESS

EDITED BY
Hang Cheng,
Fuzhou University, China

REVIEWED BY
Sinem Aslan,
University of Milan, Italy

*CORRESPONDENCE
Oshen Geenath
o.arul-jeganathan@rgu.ac.uk

RECEIVED 11 May 2025
ACCEPTED 26 August 2025
PUBLISHED 18 September 2025

CITATION
Geenath O and Priyadarshana YHPP (2025) From shades to vibrance: a comprehensive review of modern image colorization techniques. *Front. Comput. Sci.* 7:1626641. doi: 10.3389/fcomp.2025.1626641

From shades to vibrance: a comprehensive review of modern image colorization techniques

Oshen Geenath[1]* and Y. H. P. P. Priyadarshana[2]

[1]School of Computing, Robert Gordon University, Aberdeen, United Kingdom, [2]Kyoto University of Advanced Science (KUAS), Kyoto, Japan

Image colorization has become a significant task in computer vision, addressing the challenge of transforming grayscale images into realistic, vibrant color outputs. Recent advancements leverage deep learning techniques, ranging from generative adversarial networks (GANs) to diffusion models, and integrate semantic understanding, multi-scale features, and user-guided controls. This review explores state-of-the-art methodologies, highlighting innovative components such as semantic class distribution learning, bidirectional temporal fusion, and instance-aware frameworks. Evaluation metrics, including PSNR, FID, and task-specific measures, ensure a comprehensive assessment of performance. Despite remarkable progress, challenges like multimodal uncertainty, computational cost, and generalization remain. This paper provides a thorough analysis of existing approaches, offering insights into their contributions, limitations, and future directions in automated image colorization.

KEYWORDS

image colorization, real-time colorization, black-and-white colorization, user-guided colorization, interactive image colorization

1 Introduction

Image colorization, a significant task in computer vision, involves converting grayscale images into realistic and semantically consistent color outputs (Welsh et al., 2002). This technology has broad applications in historical photo restoration, film and content enhancement, digital art, and interactive media creation (Cheng et al., 2015). Despite considerable advancements, colorization remains inherently ambiguous—grayscale images may have multiple plausible colorizations depending on object semantics, scene context, and user intent (Zhang et al., 2016). Producing visually convincing results requires models to reason over both local textures and global semantic cues while maintaining computational efficiency and adaptability.

This review provides a comprehensive analysis of recent developments in deep learning-based image colorization. A systematic selection of research papers was conducted across major academic sources including Google Scholar, IEEE Xplore, ACM Digital Library, arXiv, and SpringerLink. Using search terms such as "image colorization," "automatic colorization," "semantic colorization," "user-guided colorization," and "text-to-image colorization," we identified 46 relevant publications. After excluding unrelated works on sketch-based colorization, underwater image enhancement, and extremely low-resolution inputs (Pramanick et al., 2024; Sangkloy et al., 2017; Liu et al., 2024; Lee et al., 2020; Isola et al., 2017; Fei et al., 2023; Gao et al., 2023; Saharia et al., 2022; Kumar et al., 2021; Shafiq et al., 2025; Larsson et al., 2016; Li et al., 2023), 21 influential papers published between 2015 and 2025 were selected for in-depth review.

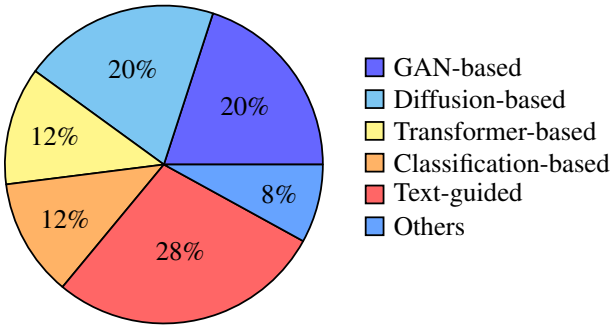

FIGURE 1
Distribution of surveyed image colorization models by architecture type (2015–2025).

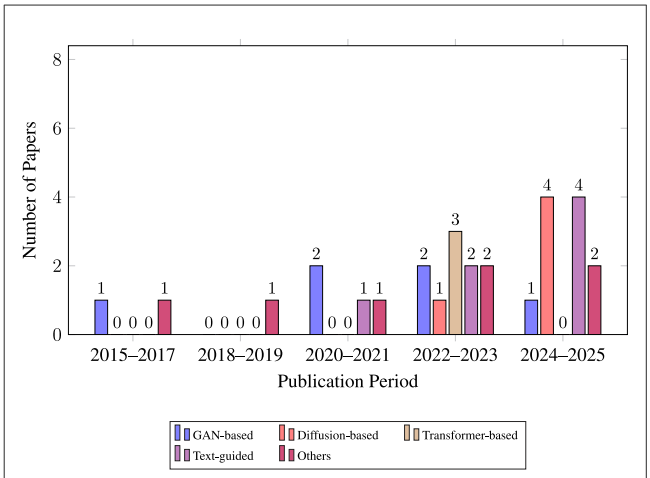

FIGURE 2
Timeline of surveyed model categories from 2015 to 2025.

This paper categorizes and evaluates state-of-the-art methodologies across seven core areas: classification-based models, adversarial networks, diffusion models, transformer and dual-decoder architectures, exemplar-based and temporal colorization, multimodal and text-guided systems, and semantic fusion-based frameworks. Each method is discussed in terms of its architectural design, innovation, quantitative performance, and limitations. In addition, a summary of benchmark datasets and widely used evaluation metrics—including PSNR, SSIM, LPIPS, FID, and CLIP Score—is provided.

The remainder of this paper is structured as follows: Section II reviews existing methods grouped by model type and design strategy. Section III outlines the key challenges facing colorization models, including color imbalance, semantic ambiguity, and computational cost. Section IV discusses evaluation metrics used to assess fidelity, diversity, and perceptual quality. Section V highlights emerging trends and future research directions, including interactive frameworks, hybrid modeling, and lightweight architectures. Section VI concludes with a summary of progress and recommendations for future research directions.

2 Existing approaches

Recent advances in image colorization have led to a diverse array of deep learning-based models that vary significantly in architectural design, learning objectives, and user controllability. This section categorizes and reviews state-of-the-art techniques into key methodological families, each offering distinct advantages and trade-offs. We organize these approaches into discretized classification models, adversarial networks, diffusion-based frameworks, transformer and dual-decoder architectures, exemplar and temporal methods, text-guided and multimodal systems, and semantic fusion models.

To enhance accessibility, we also provide a visual summary of the distribution of surveyed models by architecture type in Figure 1, highlighting how the field has evolved in terms of complexity, controllability, and realism over the past decade. This high-level overview contextualizes the detailed analysis in the subsequent subsections.

As illustrated in Figures 1, 2, GAN-based and classification-based models have historically dominated the field, while diffusion and text-guided methods have gained significant traction in recent years due to their controllability and realism. Transformer-based and multimodal approaches are also emerging, reflecting a growing emphasis on semantic alignment and user interactivity. In the following subsections, we explore each category in detail, analyzing architectural innovations, quantitative results, use cases, and limitations. A comparative summary of key image colorization models, their reported metrics, strengths, and limitations is presented in Table 1.

2.1 Discretized classification models

Regression-based colorization often results in desaturated or averaged outputs, particularly in regions with multiple plausible colors. To address this, classification-based models predict a probability distribution over discretized color classes, enhancing color diversity and rare color representation while reducing mode collapse.

Deep Colorization (Cheng et al., 2015) adopts a fully connected neural network to classify each pixel using low-, mid-, and high-level features (grayscale patches, DAISY descriptors, and semantic segmentation). While it delivers strong PSNR (up to 33 dB) and avoids CNN overhead, the lack of spatial feature reuse limits its scalability to high-resolution or texture-rich images.

Tassin et al. (2025) introduce Crayon (Tassin et al., 2025), a U-Net-based model that addresses colorization from a compression perspective using a discretized color grid. Instead of predicting full color, it reconstructs chrominance from sparse color patches retained at fixed intervals (e.g., every n^{th} pixel). This structured sampling aligns with discretized classification principles, learning color mappings from partial ground-truth. Crayon performs competitively in PSNR and CSIM across varying grid sizes ($n = 6$–100), with optimal trade-offs at $n = 15$–20. While lightweight and compression-efficient, its performance degrades at extreme sparsity levels due to color loss and grid artifacts.

TABLE 1 Overview of key image colorization models with reported metrics, strengths, and limitations.

Paper	Model type	Dataset(s)	Metrics reported	Strengths	Limitations
L-CAD Weng et al., 2023	Diffusion	COCO-Stuff, ImageNet	PSNR, SSIM	Handles fine text prompts	Prompt-sensitive, slow
SS-CycleGAN Li et al., 2023	GAN + attention	COCO	PSNR, SSIM	Spatial consistency	No FID/LPIPS, heavy
DDColor Kang et al., 2023	Transformer + CNN	ImageNet	FID, PSNR	Semantic color separation	Fails on translucent regions
L-Colns Chang et al., 2023	Transformer-based, text-guided	Extended COCO-stuff	PSNR, SSIM, LPIPS	Instance-aware without external priors	Struggles with small-object grounding in long captions
L-CoDer Chang et al., 2022	Transformer-based, text-guided	Extended COCO-stuff	PSNR, SSIM, LPIPS	Handles color-object mismatch with decoupling	High GPU/memory cost for high-resolution
L-CoDe Weng et al., 2022b	GAN-based, text-guided	COCO-stuff	PSNR, SSIM, LPIPS	High subjective realism	Color bleeding on fine boundaries
CT2 Weng et al., 2022a	Transformer-based, classification	ImageNet	PSNR, SSIM, LPIPS	Color tokens enable semantic consistency	Sensitive to biased training data
ParaColorizer Kumar et al., 2024	Dual GANs	Oxford flowers	FID, SSIM	Fast inference	Needs more training data
GAN Colorization Nazeri et al., 2018	Conditional GAN	Various	N/A	Structured training, vivid colors	Texture miscoloring
User-Guided Zhang et al., 2017	CNN + Hints	COCO	User study	Interactive and intuitive	Over-optimistic coloring
TextIR Bai et al., 2025	GAN + CLIP	CelebA, COCO	FID, SSIM, CLIP	Text-based edits	CLIP mismatch possible
Let There Be Color Welsh et al., 2002	CNN	Classic scenes	N/A	Simple, no user input needed	Fails on unseen domains
Palette Saharia et al., 2022	Diffusion	ImageNet	FID	General purpose, realistic	Slower than GANs
BiSTNet Yang et al., 2024	Video colorization (fusion)	DAVIS, Videvo	PSNR, CDC	Video accuracy, temporal logic	Heavy modules (SAM, RAFT)
Deep Colorization Cheng et al., 2015	DNN + semantic features	SUN	N/A	Few artifacts	Needs large training set
Instance-Aware Su et al., 2020	GAN + segmentation	Custom	FID	Good for multiple objects	Detection accuracy critical
ChromaGAN Vitoria et al., 2020	GAN + semantic estimation	ImageNet	PSNR	Vivid color, semantic realism	Needs labeled classes

In summary, classification-based models offer a structured way to encode color diversity and handle multimodal color spaces. They are effective for vibrant and data-driven colorization but face challenges in scalability and generalization to complex scenes due to discretization and post-processing dependencies.

2.2 Adversarial colorization networks

Generative Adversarial Networks (GANs; Fei et al., 2023) have become a cornerstone of modern image colorization, capable of producing vivid and realistic outputs by learning from natural color distributions. Unlike regression-based models, GANs use a discriminator to guide the generator toward perceptually convincing results. Recent approaches enhance this setup with semantic priors, instance awareness, and spatial refinement to boost realism and structure.

ChromaGAN (Vitoria et al., 2020) introduces a dual-branch generator: one predicts chrominance channels, the other estimates semantic class distributions, supervised by KL divergence against VGG-16 predictions. This improves contextual accuracy and color diversity. However, its reliance on fixed-size inputs (due to VGG-16 constraints) and pretrained semantic classifiers limits its adaptability across different domains, resolutions, and tasks where pretrained priors may not align with target data distributions.

ParaColorizer (Kumar et al., 2024) tackles foreground-background confusion using two parallel GANs for foreground (self-attention ResUNet) and background, fused via a DenseFuse network. This enhances object separation and color clarity, achieving top FID and colorfulness scores on COCO and ImageNet.

Its trade-off is increased complexity and inference time, along with dependency on instance segmentation.

SS-CycleGAN extends CycleGAN (Li et al., 2023) with Multi-Scale Cascaded Dilated Convolution (MCDC) and a self-attention patch discriminator, improving semantic focus and edge fidelity. It boosts PSNR and SSIM, but the model was not evaluated on perceptual realism metrics such as FID or LPIPS, limiting direct comparability with diffusion or multimodal models. Furthermore, it lacks user guidance features, reducing controllability in interactive settings.

L-CoDe (Language-based Colorization with Decoupled Conditions; Weng et al., 2022b) integrates adversarial learning with semantic decoupling by separating caption tokens into object (noun) and color (adjective) vectors, addressing color-object mismatch and coupling. A novel Attention Transfer Module (ATM) maps object references in the image to corresponding color tokens, while a Soft-gated Injection Module (SIM) ensures that only mentioned regions receive injected color guidance. The model is trained with both perceptual and binary cross-entropy losses, achieving strong performance in PSNR, SSIM, and LPIPS on the COCO-Stuff dataset. Although not evaluated on FID, L-CoDe's user studies demonstrate strong subjective realism and controllability, positioning it as a semantically guided adversarial model that bridges linguistic cues and visual fidelity.

Instance-Aware GANs (Su et al., 2020) colorize detected objects individually and merge them with global features via a fusion module, reducing color mixing between objects and backgrounds. While effective in dense scenes, the approach is highly dependent on segmentation accuracy and incurs considerable computational cost due to per-instance forward passes.

In summary, adversarial colorization networks push the boundary of realism through semantic fusion and structural refinement. Their key limitations include training complexity, runtime cost, and sensitivity to external dependencies such as detection quality and pretrained priors.

2.3 Diffusion-based colorization models

Diffusion-based models have emerged as a powerful solution for high-fidelity colorization by iteratively denoising noisy samples conditioned on grayscale input or auxiliary signals. Compared to GANs, they offer more stable training and generate diverse, semantically coherent outputs, though they remain computationally expensive and slower to infer due to their iterative nature.

Palette (Saharia et al., 2022) is a general-purpose diffusion model trained on multiple image-to-image tasks, including colorization. It uses a U-Net with global self-attention and requires no task-specific tuning. Achieving FID = 15.78 and a 47.8% human fooling rate on ImageNet, Palette outperforms earlier GAN-based models like ColTran. However, its universal design slightly compromises colorization-specific precision, and its multi-step generation makes it unsuitable for real-time use.

L-CAD (Weng et al., 2023) offers text-conditioned colorization using Stable Diffusion, integrating LIC, CEC, and ISS modules for structure preservation, semantic alignment, and object-aware control. It performs well on COCO-Stuff and ImageNet (PSNR = 26.3, SSIM = 0.911) and supports prompts of varying detail. However, its effectiveness relies on the clarity and precision of user prompts, making it vulnerable to ambiguous or sparse descriptions.

In summary, diffusion models offer high-quality, controllable colorization across modalities but face challenges in efficiency, making them ideal for offline or batch processing rather than real-time tasks. Future work must focus on faster sampling strategies and task-specific tuning to unlock their full potential in practical settings.

2.4 Transformer and dual-decoder architectures

Transformer-based and dual-decoder models have recently advanced colorization by decoupling spatial detail from semantic reasoning. This architectural split allows networks to simultaneously handle texture reconstruction and context-aware color prediction, improving accuracy in complex scenes. However, these designs often come with high training costs and memory demands.

DDColor (Kang et al., 2023) exemplifies this trend with a ConvNeXt backbone and two decoders: a pixel decoder for spatial fidelity and a transformer-based color decoder for semantic-aware color queries. Their fusion via attention mechanisms enables high-resolution, vibrant outputs. The architectural overview of DDColor is shown in Figure 3. DDColor achieves strong performance (FID = 3.92) on ImageNet, COCO-Stuff, and ADE20K, aided by a colorfulness loss. However, its dual-path design increases latency and memory usage, limiting real-time usability.

CT2 (Weng et al., 2022a) further expands transformer-based colorization by introducing color tokens and treating colorization as a classification problem in quantized color space. The model features a ViT-based encoder and a transformer-based decoder, enhanced by two novel modules: (1) a luminance-selecting module that dynamically restricts valid color candidates based on luminance levels, and (2) a color attention mechanism that injects color tokens into grayscale image features. These innovations address common issues like semantic color errors and undersaturation, leading to visually rich, plausible outputs without relying on external priors. CT2 achieves state-of-the-art performance across multiple benchmarks, including ImageNet, with superior FID (5.51), PSNR (23.50), SSIM (0.92), and colorfulness metrics. Despite its strengths, the model depends on accurate empirical color distributions and may underperform on highly biased or limited datasets.

Overall, these architectures demonstrate that semantic disentanglement improves interpretability and realism in colorization. Their main limitations are computational efficiency and generalization, which remain key areas for further refinement.

2.5 Exemplar and temporal colorization

Video colorization poses challenges like temporal consistency, color propagation, and scene coherence, which static models do not

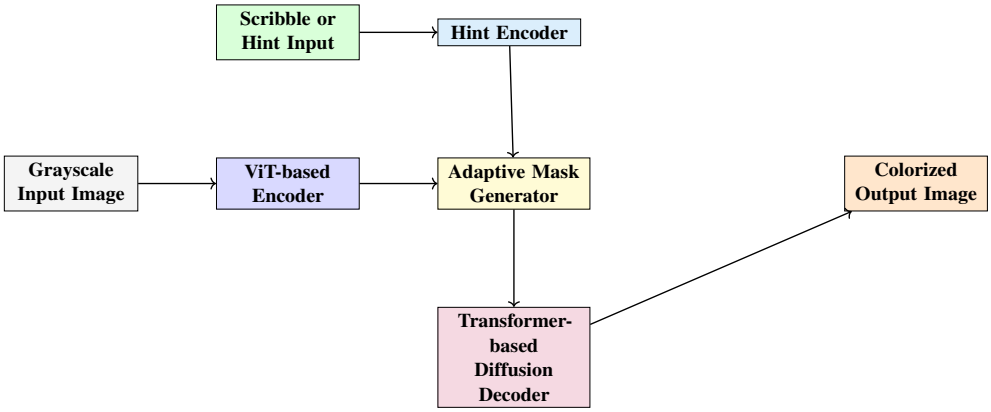

FIGURE 3
DDColor architecture: a grayscale image is encoded via a ViT encoder. Optional user hints are processed by a separate encoder and fused via an adaptive mask. A transformer-based diffusion decoder generates the final colorized image.

face. To overcome these, exemplar-based and temporal models use reference frames, semantic priors, and feature-level alignment to maintain consistency across sequences.

L-CoDer (Language-Based Colorization with Color-Object Decoupling Transformer; Chang et al., 2022) introduces a language-guided approach that unifies grayscale images and textual captions in a shared token-based representation using transformers. Unlike temporal or exemplar-based models, L-CoDer targets the modality alignment problem by decoupling the caption into noun (object) and adjective (color) tokens and processing them alongside image patches. The model employs a decoupling transformer with bidirectional attention, enabling each modality to refine the other from coarse to fine. A learned Object-Color Correspondence Matrix (OCCM) ensures correct color-object associations, addressing issues such as color-object mismatch and coupling. L-CoDer achieves state-of-the-art performance on the COCO-Stuff dataset across PSNR, SSIM, and LPIPS metrics. However, the model's transformer backbone leads to high memory demands, posing challenges for scaling to high-resolution or real-time applications. Nonetheless, it represents a strong advancement in semantically controllable colorization.

BiSTNet (Yang et al., 2024) colorizes entire video sequences using only two reference frames, employing a Bidirectional Temporal Fusion Block (BTFB) to blend forward and backward predictions based on temporal distance. It further refines output using a Mixed Expert Block (MEB)—which combines segmentation and edge features—and a Multi-Scale Refinement Block (MSRB). It achieved top scores in the NTIRE 2023 Video Colorization Challenge, with strong PSNR and CDC metrics. However, its dependency on external modules (e.g., SAM, RAFT) and heavy computation limits real-time use, and its success hinges on high-quality references.

DeepExemplar (from ParaColorizer; Kumar et al., 2024) uses a dual-GAN strategy to colorize foreground and background separately, extending it to videos through semantic alignment and temporal fusion. It preserves color consistency in repeating or structured elements and uses instance-aware segmentation for identity tracking. Despite improved visual coherence, the model remains computationally intensive, and performance degrades when semantic matching fails in dynamic or complex scenes.

In summary, these models represent a shift toward sequence-aware colorization, offering robust performance through semantic fusion and temporal logic. Their major limitations lie in runtime overhead, reference dependency, and scalability, especially in real-time or unconstrained settings.

2.6 Text-guided and multimodal colorization

Modern colorization models increasingly support multimodal interaction, enabling users to guide outputs via text prompts, strokes, or exemplars. These systems incorporate semantic understanding and visual alignment, offering both global scene-level control and localized refinement. This shift toward human-in-the-loop generation enhances creativity and personalization, but also introduces challenges in precision and usability.

TextIR (Bai et al., 2025) uses CLIP-based embeddings and a StyleConv generator to enable prompt-driven colorization, inpainting, and super-resolution. A feature fusion module blends semantic cues with structural details for fine-grained edits (e.g., "a red umbrella and green boots"). It outperforms prior models like L-CoDe and L-CoIns on FID, SSIM, and CLIP Score. However, its sensitivity to prompt quality can cause mismatches in complex or ambiguous scenarios, and its alignment is less precise than pixel-based control.

L-CAD (Weng et al., 2023) builds on Stable Diffusion, introducing modules such as LIC, CEC, and ISS to align text inputs with grayscale structure and instance features. It performs well on COCO-Stuff and ImageNet, maintaining high PSNR and SSIM, and handles both general and detailed prompts. Still, its performance heavily depends on prompt clarity, especially in multi-object scenes where ambiguity can degrade results.

Language-based Image Colorization (Li et al., 2025), a distilled diffusion model for text-guided colorization that achieves 14× faster inference and high CLIP alignment. They benchmark

from-scratch and pre-trained models, proposing a hue-invariant FID (hFID) metric for fairer evaluation. While efficient and generalizable, Color-Turbo lacks fine-grained control and may produce hue inconsistencies in complex prompts. Their curated dataset standardizes evaluation across language-based colorization models.

Controllable Image Colorization with Instance-aware Texts and Masks (An et al., 2025) extends text-based control with segmentation masks for instance-aware colorization. It combines a transformer-guided diffusion model with a novel GPT-generated dataset (GPT-Color) to enable fine-grained, object-level control. This achieves strong performance in user studies and CLIP alignment, but its reliance on accurate instance masks and multi-modal inputs increases system complexity and limits scalability for casual users.

L-CoIns (Chang et al., 2023) introduces a framework that leverages both language and instance-level cues for object-aware colorization. The model uses CLIP-based embeddings to encode text prompts and aligns them with grayscale image regions through object detection and instance segmentation. By doing so, it enables controllable, region-specific colorization (e.g., "make the apple green and the car red"). Experimental results show that L-CoIns achieves better semantic alignment and diversity compared to earlier text-based methods. However, the models effectiveness depends on accurate instance segmentation and is less responsive in cases where object boundaries are unclear or ambiguous.

In summary, text-guided and multimodal models offer flexible, user-controllable colorization, blending visual reasoning with language and manual input. Their limitations stem from prompt sensitivity, runtime demands, and precision trade-offs, but they represent a crucial step toward interactive and expressive colorization.

2.7 Semantic fusion and context-aware models

Semantic fusion models combine global scene understanding with local spatial features to guide colorization more effectively, especially in cluttered or ambiguous scenes. Through classification, segmentation, or feature alignment, these models bridge low-level texture and high-level context, resulting in more coherent and object-aware outputs.

Iizuka et al. introduced a dual-branch network with a scene classification head and a mid-level feature extractor, fused to guide per-pixel color prediction. It achieves strong perceptual realism, but lacks multimodal control and produces less diverse colors in ambiguous scenes.

ChromaGAN (Vitoria et al., 2020) operates within a GAN framework, combining color prediction with semantic class distribution estimation, regularized via KL divergence against VGG-16 outputs. This enhances realism and alignment, though the reliance on pretrained classification priors limits adaptability to unseen domains or tasks.

Instance-Aware Colorization (Su et al., 2020) improves fusion by separating object-level and global features, using Mask R-CNN (He et al., 2017) for instance detection and a fusion module to merge them. This approach excels in multi-object scenes but depends heavily on detection accuracy and incurs high computational cost when many instances are present.

BiSTNet (Yang et al., 2024), while designed for video, incorporates semantic fusion through a Mixed Expert Block (MEB) that combines segmentation and edge cues to guide color blending across frames. It achieves top performance (CDC, PSNR) but suffers from high latency due to reliance on external modules like RAFT and SAM.

In summary, semantic fusion models boost colorization accuracy by aligning structural and contextual information. Their key challenges lie in external dependencies and complexity, suggesting a need for more lightweight, integrated solutions for broader applicability.

2.8 Benchmark datasets used in colorization research

A wide range of datasets have been employed in colorization research to evaluate model performance across domains such as natural scenes, objects, faces, and videos. These datasets vary in scale, diversity, annotation detail, and complexity, enabling benchmarking on both qualitative and quantitative metrics like PSNR, SSIM, LPIPS, FID, and perceptual user studies.

ImageNet (ILSVRC2012 / val5k; Deng et al., 2009) is a large-scale dataset containing over 1.2 million labeled images across 1,000 categories. It is widely used for both training and evaluation in automatic colorization due to its semantic richness and variety of scenes. The val5k subset is a common benchmark for computing FID, PSNR, and SSIM, particularly in general-purpose and diffusion-based colorization models.

COCO-Stuff and COCO-2017 (Lin et al., 2014) datasets provide densely annotated scenes with instance-level and semantic segmentation, making them suitable for testing models like L-CAD (Weng et al., 2023)

Places205 and Places365 (Zhou et al., 2017) are scene-centric datasets with millions of labeled images across a wide range of indoor and outdoor settings. These datasets are used to support global semantic understanding, especially in models such as Iizuka et al., which incorporate scene classification into the colorization pipeline for improved contextual coherence.

CelebA and CelebA-HQ (Zhang et al., 2020) are high-quality facial datasets with attribute annotations and aligned facial landmarks, often used for portrait colorization and identity preservation. These datasets serve as testbeds for frameworks like TextIR (Bai et al., 2025) and BiSTNet (Yang et al., 2024) that require localized control or fine-grained detail in human subjects.

DAVIS and Videvo are video datasets commonly used to benchmark temporal colorization models such as BiSTNet (Yang et al., 2024). Their annotated sequences and high visual fidelity make them ideal for evaluating flicker reduction, temporal consistency, and long-range coherence in video-based colorization tasks.

Oxford 102 Flower is frequently used in models like SS-CycleGAN (Li et al., 2023) and ParaColorizer (Kumar et al., 2024) to test colorization in fine-grained textures and natural object structures. The dataset's high intra-class variance and boundary complexity help assess a models ability to retain detail.

Lastly, the SUN dataset, though smaller, is historically significant for early deep learning colorization models like Deep Colorization (Cheng et al., 2015) providing a diverse but manageable benchmark for scene understanding and category-driven colorization tasks.

3 Challenges

Despite significant progress in deep learning-based image colorization, several persistent challenges hinder model robustness, generalization, and deployment efficiency. These limitations often stem from architectural design decisions, training constraints, and dataset biases. Understanding the technical causes behind these issues is essential for improving current models and designing future systems.

One major challenge is feature imbalance in the color distribution, where dominant tones—such as grays, browns, and skin-like hues—are overrepresented in training data. This skews model predictions toward frequent colors, resulting in desaturated or uniform outputs, particularly in underrepresented regions. Classification-based models attempt to mitigate this using class reweighting and color frequency adjustment, assigning higher loss weights to rare colors. While effective, these techniques rely heavily on empirical tuning of hyperparameters, such as balancing weights and color bin definitions, which can limit generalization across datasets with different statistical properties.

Generative Adversarial Networks (GANs) present another well-known challenge: mode collapse, where the generator learns to produce a narrow set of colorizations regardless of the input diversity. This often arises from imbalanced adversarial training, where the discriminator becomes too strong and overfits to a small set of outputs, preventing the generator from exploring diverse mappings. Architectural solutions such as Wasserstein loss, spectral normalization, gradient penalties, and mini-batch discrimination have been proposed to stabilize training and encourage output diversity (Arjovsky et al., 2017; Goodfellow et al., 2020). However, these strategies often come with high computational overhead and are sensitive to training dynamics and architecture-specific constraints, making them difficult to generalize across domains or models without extensive tuning.

Semantic and spatial inconsistencies pose a significant problem, particularly in cluttered scenes with overlapping objects or ambiguous visual cues. For example, Conditional CycleGAN employs cycle-consistency loss to enforce structure preservation, but its deterministic one-to-one mapping cannot account for multi-modal color possibilities, such as a shirt that could plausibly be red or blue. As a result, these models often default to the most statistically probable color, reducing realism. Models like SS-CycleGAN improve upon this with Multi-Scale Cascaded Dilated Convolutions (MCDC) and self-attention, which expand receptive fields and allow the model to align features across spatial hierarchies (Li et al., 2023). Still, without a probabilistic mechanism, these models remain brittle in scenes with semantic ambiguity. In contrast, VAEs and diffusion models incorporate stochastic sampling and latent-variable conditioning, making them better suited for uncertainty modeling and diverse color prediction—but often at the expense of inference speed and simplicity.

Another widespread issue is structural distortion, including edge noise, color bleeding, and boundary mismatch. These problems are especially evident in models without strong instance-awareness or edge supervision. Recent models like CtrlColor integrate SAM-based segmentation and edge-aware loss functions to preserve object boundaries. While these methods improve sharpness and local consistency, they often rely on external modules (e.g., SAM or RAFT) and high-resolution computation, which increase runtime complexity and limit real-time applicability on low-power devices.

In summary, image colorization remains a multi-dimensional optimization problem. Models must balance color diversity, semantic fidelity, spatial structure, and computational efficiency. Each class of architecture addresses some of these goals but introduces new trade-offs. The path forward lies in hybrid designs that combine deterministic structure preservation with probabilistic color reasoning, along with lightweight, end-to-end architectures that minimize external dependencies while supporting interactive and real-time applications. To better understand how recent colorization models perform in real-world settings, we provide a comparative analysis in Table 2. This table summarizes the practical usability of key models based on real-time capability, inference time, and hardware requirements. Such comparisons are essential for selecting appropriate models for deployment on edge devices, real-time systems, or cloud platforms.

4 Evaluation metrics

Evaluation metrics are essential for assessing the performance and quality of image colorization methods. They provide quantitative and qualitative insights into how well a model performs, ensuring comprehensive evaluation from multiple perspectives. The metrics used in image colorization are broadly categorized into pixel-wise accuracy, structural and perceptual similarity, generative quality, and task-specific measures.

4.1 Pixel-wise accuracy

Pixel-level metrics, such as Mean Squared Error (MSE) and Peak Signal-to-Noise Ratio (PSNR), are commonly employed to evaluate the fidelity of generated images against the ground truth (Hore and Ziou, 2010). MSE measures the pixel-wise differences, ensuring accurate reconstruction at the pixel level. The formula for MSE is given as:

$$\mathrm{MSE} = \frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W}\left(X_{ij} - \hat{X}_{ij}\right)^2$$

where H and W are the height and width of the image, and X_{ij} and $\hat{X}_{ij}$ are the ground-truth and generated pixel values, respectively. PSNR, on the other hand, reflects the reconstruction quality and is computed as:

$$\mathrm{PSNR} = 20 \cdot \log_{10}\left(\frac{\mathrm{MAX}}{\sqrt{\mathrm{MSE}}}\right)$$

TABLE 2 Inference performance and real-time capability of key image colorization models.

Model	Control mode	Inference time	Hardware used	Real-time
DDColor Kang et al., 2023	None	N/A	4 Tesla V100	No
ParaColorizer Kumar et al., 2024	None	∼0.24 ms	2 Tesla V100	Yes
TextIR Bai et al., 2025	Text	N/A	2 Tesla V100	No
L-Colns Chang et al., 2023	Text	N/A	8 RTX 3090	No
L-CoDer Chang et al., 2022	Text	N/A	4 NVIDIA TITAN TRX	No
L-CoDe Weng et al., 2022b	Text	N/A	2 GTX 1080Ti	No
CT2 Weng et al., 2022a	None	N/A	8 RTX 3090	No
Palette Saharia et al., 2022	None	∼0.8 s	TPU v3	No
SS-CycleGAN Li et al., 2023	None	N/A	Tesla T4	No
L-CAD Weng et al., 2023	Text	N/A	2 RTX 3090Ti	No
Instance-Aware GAN Su et al., 2020	None	∼0.187 s	RTX 2080Ti	No
ChromaGAN Vitoria et al., 2020	None	∼4.4 ms	Quadro P6000	No
GAN Colorization Nazeri et al., 2018	None	N/A	N/A	No
User-Guided Zhang et al., 2017	User Hint	N/A	N/A	Yes
Let There Be Color Welsh et al., 2002	None	N/A	CPU	No
Deep Colorization Cheng et al., 2015	None	N/A	Tesla K40	No
BiSTNet Yang et al., 2024	Reference frames	N/A	4 RTX A6000	No

where MAX is the maximum pixel value in the image. These metrics are extensively used in methods like L-CAD, and DDColor to evaluate the accuracy of chrominance predictions (Weng et al., 2023; Kang et al., 2023). While effective, these metrics may not fully capture the perceptual quality of colorization outputs, especially in multimodal tasks.

4.2 Structural and perceptual similarity

Structural and perceptual similarity metrics are crucial for evaluating the consistency of structural and visual coherence between the generated and ground-truth images (Wang et al., 2004). The Structural Similarity Index (SSIM) measures luminance, contrast, and structural similarity using the following equation:

$$\mathrm{SSIM}(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}$$

where μ_x, μ_y are the means, σ_x^2, σ_y^2 are the variances, and σ_{xy} is the covariance of the two images, with C_1 and C_2 being constants. Learned Perceptual Image Patch Similarity (LPIPS) evaluates perceptual similarity by comparing deep feature representations (Zhang et al., 2018), as follows:

$$\mathrm{LPIPS}(x,y) = \sum_l \frac{1}{H_l W_l} \sum_{i,j} \|\phi_l(x)_{ij} - \phi_l(y)_{ij}\|_2^2$$

where $\phi_l(x)$ and $\phi_l(y)$ are features from layer l, and H_l, W_l are the dimensions of the feature map. Metrics such as SSIM and LPIPS are widely used in methods like SS-CycleGAN, ParaColorizer, and BiSTNet to ensure structural coherence and perceptual quality (Wang et al., 2004; Zhang et al., 2018; Li et al., 2023; Kumar et al., 2024; Yang et al., 2024).

4.3 Generative quality

Generative quality metrics, such as Fréchet Inception Distance (FID) and Inception Score (IS), measure the realism and diversity of generated images (Heusel et al., 2017; Barratt and Sharma, 2018). FID quantifies the similarity between the distributions of real and generated image features and is given by:

$$\mathrm{FID} = \|\mu_r - \mu_g\|^2 + \mathrm{Tr}(\Sigma_r + \Sigma_g - 2(\Sigma_r\Sigma_g)^{1/2})$$

where μ_r, μ_g are the means and Σ_r, Σ_g are the covariances of real and generated image features. Inception Score (IS) evaluates diversity and quality by analyzing the entropy of predictions made by a pre-trained classification model. Generative metrics like FID are commonly employed in methods such as Palette, and DDColor to ensure that generated outputs are both realistic and diverse (Saharia et al., 2022; Kang et al., 2023). Additionally, the Colorfulness Metric assesses the vibrancy and richness of colors in generated images, reflecting the vividness of the results.

4.4 Qualitative and perceptual evaluations

In addition to quantitative measures, qualitative evaluations and user studies play a vital role in assessing the perceptual realism of colorized images. Perceptual studies, as conducted in methods like ChromaGAN and Real-Time User-Guided Colorization, involve measuring fooling rates and user preferences (Vitoria et al., 2020; Zhang et al., 2017). These evaluations complement traditional metrics by capturing subjective qualities such as naturalness and believability, particularly in multimodal and visually ambiguous scenarios.

5 Emerging trends and future directions

5.1 Emerging trends in image colorization

5.1.1 Diffusion models as the new backbone

The success of models like Palette (Saharia et al., 2022) has led to a shift from GANs to diffusion models for high-fidelity, controllable colorization. Diffusion provides better color diversity and supports iterative refinements, making it more suitable for creative tasks. However, high inference latency remains a bottleneck, limiting real-time use.

5.1.2 Prompt-based and multimodal interaction

Prompt-guided models like L-CAD and TextIR (Weng et al., 2023; Bai et al., 2025) illustrate how text can guide colorization flexibly, even at the region level. With increasing adoption of CLIP and similar models, the future may lean toward foundation model-guided colorization, allowing zero-shot or few-shot customization using natural language.

5.1.3 Real-time and lightweight inference

Models like ParaColorizer (Kumar et al., 2024) and User-Guided Colorization (Zhang et al., 2017) reflect an increasing demand for real-time colorization, particularly for mobile and AR/VR applications. Future systems will likely prioritize architectures that trade off minimal quality for fast, efficient deployment on edge devices.

5.1.4 Ethics, bias mitigation, and explainability

As image colorization systems move beyond artistic applications and into sensitive domains—such as historical restoration, forensic analysis, and medical imaging—the need for ethical safeguards has become increasingly urgent. A primary ethical challenge is the risk of color misinterpretation. When models hallucinate colors without ground truth references, they may inadvertently introduce misleading or historically inaccurate information. For example, assigning skin tones or fabric colors in archival photographs could distort cultural or racial identity, leading to unintentional misrepresentation of the past.

Another major concern is dataset bias. Popular datasets such as COCO or ImageNet often reflect implicit social and cultural biases, which can propagate into generated outputs. This may result in systematically skewed colorizations—for instance, consistently rendering certain demographics with particular tones—thereby reinforcing stereotypes or marginalizing underrepresented groups.

In high-stakes domains like journalism or forensics, hallucinated colorizations may be mistaken as factual, particularly when presented without proper disclaimers. In evidentiary settings, such misinterpretations could even carry legal implications. This underscores the importance of embedding uncertainty visualization, provenance tracking, and clear disclaimers to differentiate generated content from original data.

Explainability also remains limited. While some recent systems integrate user hints, segmentation cues, or attention mechanisms to guide outputs, most colorization pipelines remain opaque to end users. This lack of transparency hinders trust and accountability, especially in workflows where factual accuracy is paramount.

To mitigate these concerns, future research should emphasize transparency mechanisms such as attribution maps, error bounds, and dataset audits. Additionally, incorporating controllable generation frameworks with provenance logging can empower users to better understand and guide the colorization process—promoting both ethical integrity and user trust.

5.2 Future research directions

Emerging trends in image colorization highlight the shift toward hybrid transformer-convolutional architectures, structurally-aware learning, prompt-driven multimodal control, and real-time interactivity. While current models achieve photorealism and semantic richness, challenges remain in scalability, boundary preservation, and global reasoning.

A notable direction is the move from traditional CNN backbones (e.g., VGG-16 in ChromaGAN) to hybrid architectures combining ConvNeXt and Transformers (Vitoria et al., 2020; Kang et al., 2023; Liu et al., 2021). Models like DDColor show how ConvNeXt can preserve textures while transformers enhance contextual reasoning (Liu et al., 2022). A hybrid multi-scale architecture with structured attention fusion could improve local-global feature integration, addressing issues like over-smoothing and poor generalization.

Color bleeding remains a challenge in GAN-based models (Li et al., 2023; Kumar et al., 2024; Nazeri et al., 2018). Recent solutions propose edge-conditioned discriminators (e.g., using Canny or HED maps) and boundary-aware generators with edge-guided attention. A dual-weighted loss that balances perceptual smoothness with structural sharpness could further improve fidelity and boundary accuracy.

Multimodal frameworks such as L-CAD and TextIR reflect another key trend, enabling prompt-guided and user-controllable colorization via text or exemplars (Weng et al., 2023; Bai et al., 2025). These systems offer customization and interactivity, paving the way for integration into creative tools and restoration pipelines.

In summary, the future of colorization lies in developing interactive, scalable, and semantically aligned systems. Through architectural innovation and user-focused design, next-generation models will support applications in digital media, heritage restoration, and augmented creativity.

6 Conclusion

Image colorization has advanced significantly, driven by deep learning, semantic understanding, and generative models. This review explored innovations such as semantic class distributions, multimodal fusion, and user-guided controls, addressing challenges like multimodal uncertainty and object-level consistency. Despite these advancements, limitations such as high computational costs, dataset dependencies, and performance on unseen scenarios remain. Future work should focus on lightweight models, enhanced generalization, and interactive frameworks to balance automation with creative flexibility. Transforming grayscale to vibrant color continues to be an exciting frontier in computer vision.

Author contributions

OG: Formal analysis, Investigation, Validation, Writing – original draft, Writing – review & editing. YP: Supervision, Writing – review & editing.

Funding

The author(s) declare that no financial support was received for the research and/or publication of this article.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Gen AI was used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

References

An, Y., Gui, L., Hu, Q., Cai, C., Ye, T., Zhang, X., et al. (2025). Controllable image colorization with instance-aware texts and masks. *arXiv preprint arXiv:2505.08705*. doi: 10.48550/arXiv.2505.08705

Arjovsky, M., Chintala, S., and Bottou, L. (2017). "Wasserstein generative adversarial networks," in *International Conference on Machine Learning (ICML)* (Sydney, NSW: MLR), 214–223.

Bai, Y., Wang, C., Xie, S., Dong, C., Yuan, C., and Wang, Z. (2025). "Textir: a simple framework for text-based editable image restoration," in *IEEE Transactions on Visualization and Computer Graphics* (Piscataway, NJ: IEEE). doi: 10.1109/TVCG.2025.3550844

Barratt, S., and Sharma, R. (2018). A note on the inception score. *arXiv preprint arXiv:1801.01973*. doi: 10.48550/arXiv.1801.01973

Chang, Z., Weng, S., Li, Y., Li, S., and Shi, B. (2022). "L-coder: language-based colorization with color-object decoupling transformer," in *European Conference on Computer Vision* (Springer: New York), 360–375. doi: 10.1007/978-3-031-19797-0_21

Chang, Z., Weng, S., Zhang, P., Li, Y., Li, S., and Shi, B. (2023). "L-coins: language-based colorization with instance awareness," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Vancouver, BC: IEEE), 19221–19230. doi: 10.1109/CVPR52729.2023.01842

Cheng, Z., Yang, Q., and Sheng, B. (2015). "Deep colorization," in *Proceedings of the IEEE International Conference on Computer Vision* (Santiago: IEEE), 415–423. doi: 10.1109/ICCV.2015.55

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). "Imagenet: a large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition* (Miami: IEEE), 248–255. doi: 10.1109/CVPR.2009.5206848

Fei, B., Lyu, Z., Pan, L., Zhang, J., Yang, W., Luo, T., et al. (2023). "Generative diffusion prior for unified image restoration and enhancement," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Vancouver, BC: IEEE), 9935–9946. doi: 10.1109/CVPR52729.2023.00958

Gao, X., Mou, J., Banerjee, S., and Zhang, Y. (2023). Color-gray multi-image hybrid compression-encryption scheme based on bp neural network and knight tour. *IEEE Trans. Cybern.* 53, 5037–5047. doi: 10.1109/TCYB.2023.3267785

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2020). Generative adversarial networks. *Commun. ACM* 63, 139–144. doi: 10.1145/3422622

He, K., Gkioxari, G., Dollár, P., and Girshick, R. (2017). "Mask R-CNN," in *Proceedings of the IEEE International Conference on Computer Vision* (Venice: IEEE), 2961–2969. doi: 10.1109/ICCV.2017.322

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. (2017). Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Adv. Neural. Inf. Process Syst.* 30, 6626–6637. doi: 10.48550/arXiv.1706.08500

Hore, A., and Ziou, D. (2010). "Image quality metrics: PSNR vs. SSIM," in *2010 20th International Conference on Pattern Recognition (ICPR)* (Istanbul: IEEE), 2366–2369. doi: 10.1109/ICPR.2010.579

Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. (2017). "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Honolulu, HI: IEEE), 1125–1134. doi: 10.1109/CVPR.2017.632

Kang, X., Yang, T., Ouyang, W., Ren, P., Li, L., and Xie, X. (2023). "Ddcolor: towards photo-realistic image colorization via dual decoders," in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (Paris: IEEE), 328–338. doi: 10.1109/ICCV51070.2023.00037

Kumar, H., Banerjee, A., Saurav, S., and Singh, S. (2024). Paracolorizer-realistic image colorization using parallel generative networks. *Vis. Comput.* 40, 4039–4054. doi: 10.1007/s00371-023-03067-7

Kumar, M., Weissenborn, D., and Kalchbrenner, N. (2021). Colorization transformer. *arXiv preprint arXiv:2102.04432*. doi: 10.48550/arXiv.2102.04432

Larsson, G., Maire, M., and Shakhnarovich, G. (2016). "Learning representations for automatic colorization," in *Computer Vision-ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14* (Springer: New York), 577–593. doi: 10.1007/978-3-319-46493-0_35

Lee, J., Kim, E., Lee, Y., Kim, D., Chang, J., and Choo, J. (2020). "Reference-based sketch image colorization using augmented-self reference and dense semantic correspondence," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Seattle, WA: IEEE), 5801–5810. doi: 10.1109/CVPR42600.2020.00584

Li, B., Lu, Y., Pang, W., and Xu, H. (2023). Image colorization using cyclegan with semantic and spatial rationality. *Multimed. Tools Appl.* 82, 21641–21655. doi: 10.1007/s11042-023-14675-9

Li, Y., Yang, S., and Liu, J. (2025). Language-based image colorization: a benchmark and beyond. *arXiv preprint arXiv:2503.14974*. doi: 10.48550/arXiv.2503.14974

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., et al. (2014). "Microsoft coco: common objects in context," in *Computer Vision-ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13* (Springer: New York), 740–755. doi: 10.1007/978-3-319-10602-1_48

Liu, Y., Zhao, H., Chan, K. C., Wang, X., Loy, C. C., Qiao, Y., et al. (2024). Temporally consistent video colorization with deep feature propagation and self-regularization learning. *Comput. Vis. Media* 10, 375–395. doi: 10.1007/s41095-023-0342-8

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., et al. (2021). "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (Montreal, QC: IEEE), 10012–10022. doi: 10.1109/ICCV48922.2021.00986

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., and Xie, S. (2022). "A convnet for the 2020s," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (New Orleans, LA: IEEE), 11976–11986. doi: 10.1109/CVPR52688.2022.01167

Nazeri, K., Ng, E., and Ebrahimi, M. (2018). "Image colorization using generative adversarial networks," in *Articulated Motion and Deformable Objects: 10th International Conference, AMDO 2018, Palma de Mallorca, Spain, July 12-13, 2018, Proceedings 10* (Springer: New York), 85–94. doi: 10.1007/978-3-319-94544-6_9

Pramanick, A., Sarma, S., and Sur, A. (2024). "X-caunet: Cross-color channel attention with underwater image-enhancing transformer," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing* (Seoul: IEEE), 3550–3554. doi: 10.1109/ICASSP48485.2024.10445832

Saharia, C., Chan, W., Chang, H., Lee, C., Ho, J., Salimans, T., et al. (2022). "Palette: image-to-image diffusion models," in *ACM SIGGRAPH 2022 Conference Proceedings* (Vancouver, BC: ACM), 1–10. doi: 10.1145/3528233.3530757

Sangkloy, P., Lu, J., Fang, C., Yu, F., and Hays, J. (2017). "Scribbler: controlling deep image synthesis with sketch and color," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Honolulu, HI: IEEE), 5400–5409. doi: 10.1109/CVPR.2017.723

Shafiq, H., Nguyen, T., and Lee, B. (2025). Colorformer: a novel colorization method based on a transformer. *Neurocomputing* 649:130743. doi: 10.1016/j.neucom.2025.130743

Su, J.-W., Chu, H.-K., and Huang, J.-B. (2020). "Instance-aware image colorization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Seattle, WA: IEEE), 7968–7977. doi: 10.1109/CVPR42600.2020.00799

Tassin, I., Goebel, K., and Lasher, B. (2025). Convolutional deep colorization for image compression: a color grid based approach. *arXiv preprint arXiv:2502.05402*. doi: 10.48550/arXiv.2502.05402

Vitoria, P., Raad, L., and Ballester, C. (2020). "Chromagan: adversarial picture colorization with semantic class distribution," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (Snowmass Village, CO: IEEE), 2445–2454. doi: 10.1109/WACV45572.2020.9093389

Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.* 13, 600–612. doi: 10.1109/TIP.2003.819861

Welsh, T., Ashikhmin, M., and Mueller, K. (2002). "Transferring color to greyscale images," in *Proceedings of the 29th Annual Conference on Computer Graphics and Interactive Techniques* (San Antonio, TX: ACM), 277–280. doi: 10.1145/566570.566576

Weng, S., Sun, J., Li, Y., Li, S., and Shi, B. (2022a). "Ct 2: colorization transformer via color tokens," in *European Conference on Computer Vision* (Springer: New York), 1–16. doi: 10.1007/978-3-031-20071-7_1

Weng, S., Wu, H., Chang, Z., Tang, J., Li, S., and Shi, B. (2022b). L-code: language-based colorization using color-object decoupled conditions. *Proc. AAAI Conf. Artif. Intell.* 36, 2677–2684. doi: 10.1609/aaai.v36i3.20170

Weng, S., Zhang, P., Li, Y., Li, S., Shi, B., et al. (2023). L-cad: language-based colorization with any-level descriptions using diffusion priors. *Adv. Neural Inf. Process. Syst.* 36, 77174–77186. doi: 10.48550/arXiv.2310.14191

Yang, Y., Pan, J., Peng, Z., Du, X., Tao, Z., and Tang, J. (2024). Bistnet: semantic image prior guided bidirectional temporal feature fusion for deep exemplar-based video colorization. *IEEE Trans. Pattern Anal. Mach. Intell.* 46, 5612–5624. doi: 10.1109/TPAMI.2024.3370920

Zhang, R., Isola, P., and Efros, A. A. (2016). "Colorful image colorization," in *Computer Vision-ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14* (Springer: New York), 649–666. doi: 10.1007/978-3-319-46487-9_40

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. (2018). "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Salt Lake City, UT: IEEE), 586–595. doi: 10.1109/CVPR.2018.00068

Zhang, R., Zhu, J.-Y., Isola, P., Geng, X., Lin, A. S., Yu, T., et al. (2017). Real-time user-guided image colorization with learned deep priors. *arXiv preprint arXiv:1705.02999*. doi: 10.48550/arXiv.1705.02999

Zhang, Y., Yin, Z., Li, Y., Yin, G., Yan, J., Shao, J., et al. (2020). "Celeba-spoof: large-scale face anti-spoofing dataset with rich annotations," in *Computer Vision-ECCV 2020: 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XII 16* (Springer: New York), 70–85. doi: 10.1007/978-3-030-58610-2_5

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., and Torralba, A. (2017). Places: a 10 million image database for scene recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* 40, 1452–1464. doi: 10.1109/TPAMI.2017.2723009