From Gray to Vivid: Methods and Metrics for Image Colorization

Riya Ann Easow
EE21BTECH11044

Beaula Mahima V
MA21BTECH11002

Prasham Walvekar
CS21BTECH11047

Kallu Rithika
AI22BTECH11010

Abstract

Image colorization is a fundamental task in computer vision that aims to restore plausible and aesthetically pleasing colors to grayscale images. Recent advances in deep learning have led to the development of generative adversarial networks (GANs), convolutional neural networks (CNNs), and transformer-based approaches for colorization. Despite significant progress, evaluating colorization remains a challenge because of the inherent difficulties in gauging the perceptual and natural qualities of image colorings.

1. Introduction

Color plays a crucial role in computer vision, influencing tasks such as object detection, tracking, and recognition. Image colorization is the process of assigning realistic colors to grayscale images, a task that requires understanding object semantics, lighting conditions, and natural color distributions. Important real-world applications include restoring historical photographs, enhancing medical imaging, improving autonomous driving perception, and aiding creative content generation.

The problem is severely ill-posed as two out of the three image dimensions are missing. While humans intuitively assign colors based on their understanding of the world, automatic image colorization remains a challenging problem due to its inherent ambiguity—many objects can have multiple plausible colorizations. Early colorization methods relied on reference images or user inputs, but recent advances in deep learning have enabled fully automatic approaches.

Deep learning approaches, particularly CNNs, GANs, VAEs, and Transformers, have significantly advanced this field by leveraging large datasets and learning complex image features. However, despite these improvements, challenges remain in balancing realism, diversity, computational efficiency, and handling a wide range of object categories and scenes. Additionally, evaluating the quality of colorized images is difficult, as traditional metrics like PSNR (Peak Signal-to-Noise Ratio) and SSIM (Structural Similarity Index) fail to capture perceptual realism, while newer metrics like FID (Fréchet Inception Distance) and LPIPS (Learned Perceptual Image Patch Similarity) still have limitations.

2. Literature Review

The automatic image colorization techniques can be broadly divided into 3 categories: Early Architectures, Diverse Colorization Networks, Multi-path Networks. [1]

2.1. Early Architectures

Early colorization models used simple architectures that stacked convolutional layers with no or naive skip connections. While easy to implement, they required more training data and had slow learning.

Deep Colorization[6] proposes the first method to incorporate CNNs for the image colorization, eliminating the need for manual input by leveraging a large-scale dataset to map grayscale features to chrominance values. A joint bilateral filtering[21] based post processing step ensures artifact free results.

Colorful Image Colorization[28] proposes a fully automatic CNN based approach to colorize grayscale images into plausible, vibrant color images. It introduces a class-rebalancing technique to enhance rare colors, preventing desaturated outputs. It demonstrates the use of colorization as a pretext task for self-supervised feature learning. It also achieves state-of-the-art performance on several benchmarks for representation learning.

Deep Depth Colorization[5] presents a deep learning approach for depth image colorization, aimed at improving object recognition. This system is primarily designed for object recognition by learning the mapping from the depths to RGB channels. The $(DE)^2CO$ algorithm outperforms traditional handcrafted mappings, achieving significant performance gains.

U-Net Grayscale Image Colorization[12] proposes method using an improved U-Net CNN. The approach operates in the space of Lab color model, where L component (Lightness) is used to predict the a and b color components. The research sets the stage for further improvements in automatic colorization using more advanced architectures.

Figure 1. Colorized output images alongside their corresponding grayscale inputs.

2.2. Diverse Colorization Networks

Various deep learning-based techniques, primarily using Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs), have been developed to achieve this goal. Some approaches generate multiple diverse colorizations, while others predict a single realistic colorization based on semantic information.

Cao et al. [4] introduced unsupervised diverse colorization using conditional GANs, where noise channels were added to the generator to ensure multiple diverse outputs. Nazeri et al. [18] developed ICGAN inspired by fully convolutional networks for semantic segmentation, which predicts a single realistic colorization by taking grayscale images as input. The generator's modified cost function improved the quality of the colorized outputs compared to traditional CNN-based methods.

Frans [10] proposed Tandem Adversarial Networks for colorizing line art, predicting a single colorization. The model used two adversarial networks: one for color prediction and another for shading. Skip connections and L2 or adversarial losses improved the final output. Deshpande et al. [9] employed a VAE + Mixture Density Network (MDN) to generate multiple diverse yet realistic colorizations. The VAE learned a low-dimensional embedding, while MDN introduced diversity by sampling different plausible colorizations. The approach was enhanced by three loss functions: specificity, colorfulness, and gradient.

ChromaGAN [23] leveraged self-supervised learning and semantic understanding to predict a single realistic colorization rather than focusing on diversity. The model used two branches: one predicting chrominance information and another generating a class distribution vector. Using PatchGAN as a discriminator, ChromaGAN achieved highly realistic results when trained on 1.3M ImageNet images. MemoPainter [26] integrated memory networks with GAN-based colorization to learn from limited data and predict a single accurate colorization for objects with consistent colors. A threshold triplet loss was introduced to help the memory network store rare object colors and retrieve them for more accurate outputs.

Li et al. [16] proposed SS-CycleGAN, which used two generators and two discriminators to translate grayscale images to color and vice versa. The model predicts a single colorization but improves semantic and structural consistency using a self-attention mechanism and multi-scale cascaded dilated convolutions. Shafiq and Lee [22] developed a Transformer-GAN hybrid for colorization. The VGG-based encoder captured global semantics, while a Swin Transformer block processed color-specific features. The model can predict multiple plausible colorizations, and the GAN framework further improved the visual quality. The final output was refined using perceptual, adversarial, and color losses.

2.3. Multi-path Networks

Multi-path networks learn different features in various branches. This technique allows the model to learn an accurate representation of multiple features. However, due to the complex nature of these architectures, it is more computationally expensive.

Iizuka et al. [13] proposed 'Let there be color,' a CNN-based model that fuses global and local features for automatic image colorization. The model has four subnets, integrating scene classification with chrominance prediction through a joint colorization classification loss. Unlike previous methods, it enables resolution independent colorization by adapting colors based on image context. Further, Larsson et al. [15] proposed a deep learning model that predicts per-pixel hue and chroma distributions using hypercolumns from VGG16. The model extracts a 12k-channel feature descriptor per pixel, mapping it to color predictions through a fully connected layer with KL divergence loss. By leveraging both low-level and semantic representations, it achieves visually coherent and realistic colorization.

PixColor [11] employs a two-stage approach, using a conditional PixelCNN for low-resolution color prediction followed by a refinement CNN for high-resolution colorization. It leverages masked and gated convolutions to capture spatial dependencies. By constraining PixelCNN sampling to low-resolution chroma channels, PixColor achieves fast inference while maintaining high-quality results. ColorCapsNet [19] enhances CapsNet by integrat-

ing VGG-19 convolutional layers for feature extraction, adding batch normalization, and reducing the number of capsules. It learns color distributions in the CIE Lab space from patches and reconstructs full images while minimizing mean squared error loss.

Further, Zhao et al. introduced Pixelated [30], a colorization model that integrates pixelated semantic understanding using a dual-branch network—one for color embedding and another for semantic segmentation. It employs an autoregressive approach with a conditional Pixel-CNN for diverse colorization, while multi-scale features extracted via atrous spatial pyramid pooling enhance semantic representation.

Mohammad et al. [2] proposed a deep tree-structured network for image colorization to use in image compression, generating multiple color hypotheses per pixel instead of a single prediction. The method leverages a common convolutional trunk to extract shared features and stores minimal additional information to reconstruct true colors with high fidelity.

2.4. Evaluation Metrics

Evaluating image colorization remains an open challenge, as traditional metrics like PSNR (Peak Signal-to-Noise Ratio) and SSIM (Structural Similarity Index) often fail to capture the perceptual quality of colorized images. Consequently, a number of perceptual and naturalness evaluation metrics ave been explored. However, most image colorization models in the literature have been assessed through human evaluation, which remains the gold standard to this day. For example, in [28], a "Colorization Turing Test" was conducted, where human observers were asked to distinguish between real and colorized images. Similar subjective methods were used to evaluate model performance in [8, 13].

However, human evaluation is subjective, time-consuming, expensive, and not scalable, making it impractical for large-scale assessments. Therefore, there is a pressing need for more efficient and scalable metrics tailored specifically to the colorization task.

3. Datasets

Choosing the right dataset is crucial for training an image colorization model that can handle diverse cases effectively.

1. **COCO-Stuff** [3] contains 164k images across 172 categories, covering complex real-world scenes. Although the colors in the dataset appear unnatural due to stuff annotations, this helps the model learn to colorize different parts of an image separately.
2. **ImageNet** [7] consists of 1.2M images spanning 1000 categories, featuring varied objects, lighting conditions, and real-world diversity. This dataset enhances the model's ability to colorize a wide range of objects accurately and generalize well to new images.
3. **Places205** [31] provides 2.5M images across 205 scene categories, focusing on indoor and outdoor environments rather than individual objects. This dataset is valuable for learning scene-based colorization, ensuring the model captures natural color distributions in different locations.

Other datasets were considered but found less suitable. CIFAR-100's low resolution (32×32) is a limitation, while Pascal VOC's 11k images focus on segmentation-based colorization with limited scale and variety. Instead, we choose COCO-Stuff, Places205, and ImageNet for their complementary strengths: Places205 enhances scene understanding, while COCO-Stuff and ImageNet provide diverse, large-scale data for better real-world adaptability. Further, since we do not require any labels for our task, we could make use of the large amounts of unlabeled data available by web scraping, to model the natural color distribution more accurately.

4. Project Overview

4.1. Proposed Approach

Image colorization is a rapidly evolving field with significant room for improvement. In this project, we aim to analyze state-of-the-art architectures and explore ways to enhance their performance. One possible approach is to incorporate naturalness constraints into the loss function. For instance, modeling NIQE (Natural Image Quality Evaluator) [17] as a loss function could guide the model to generate more perceptually natural color distributions similar to the approach followed in [27] in the field of image super-resolution. Additionally, performance could be potentially improved by leveraging an ensemble of state-of-the-art architectures or introducing additional pre-processing modules.

4.2. Model Evaluation

We address the issues mentioned in 2.4 by analyzing the performance of the proposed models using an evaluation framework carefully curated for the colorization task. The framework comprises metrics and methodologies that fall into either of the following categories.

1. **Perceptual and Naturalness Quality Metrics**:
   Image colorization is inherently a perceptual task, meaning its success depends on how natural and realistic the colorized images appear to the human eye. Therefore, metrics such as LPIPS [29], which measures perceptual similarity using deep neural network features, and NIQE, a no-reference metric assessing the naturalness of an image based on statistical properties, can provide valuable insights into the perceptual quality of colorized images.

2. **Performance Comparison on Downstream Tasks**:
   An intuitive way to assess the quality of our image colorization is by analyzing how pre-trained models for downstream tasks respond to the colorized output compared to the original colored image. To perform such an evaluation, we select a suitable downstream task (such as classification) and a pre-trained model (e.g. ResNet, VGG), then compare the model's behavior on the original image versus the colorized output.
3. **Color Specific Metrics**:
   Since the task at hand is colorization, it is essential to use metrics that assess color fidelity, balance, and contrast. In this regard, we plan to explore metrics such as PCQI (Perceptual Color Image Quality) [24], UIQM (Underwater Image Quality Measure) [20], and QSSIM (Quaternion Structural Similarity Index) [14], as they are specifically designed to evaluate the quality of color images.
4. **Color Harmony Inspired Metrics and Methods**:
   Since the primary goal of image colorization in most applications is to enhance the aesthetics of gray-scale images, we believe that evaluating our model's colorization based on principles of color harmony is well-suited. This approach allows us to assess the aesthetic appeal of the colorized images.
   In our project, we plan to explore concepts such as those presented in [25] to develop metrics and methods for evaluating the aesthetics of colorized images. Since this evaluation approach has not yet been explored for the colorization task, successfully implementing it would bring forth a unique contribution to this domain.

These metrics and methods collectively account for various factors that could influence a human evaluator's decision, providing a comprehensive yet scalable assessment of the colorization.

References

[1] Saeed Anwar, Muhammad Tahir, Chongyi Li, Ajmal Mian, Fahad Shahbaz Khan, and Abdul Wahab Muzaffar. Image colorization: A survey and dataset. *Information Fusion*, 114: 102720, 2025. 1

[2] Mohammad Haris Baig and Lorenzo Torresani. Multiple hypothesis colorization and its application to image compression. *Computer Vision and Image Understanding*, 164:111–123, 2017. 3

[3] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Cocostuff: Thing and stuff classes in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1209–1218, 2018. 3

[4] Yun Cao, Zhiming Zhou, Weinan Zhang, and Yong Yu. Unsupervised diverse colorization via generative adversarial networks. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2017, Skopje, Macedonia, September 18–22, 2017, Proceedings, Part I 10*, pages 151–166. Springer, 2017. 2

[5] Fabio Maria Carlucci, Paolo Russo, and Barbara Caputo. ‘2 co: Deep depth colorization. *IEEE Robotics and Automation Letters*, 3(3):2386–2393, 2018. 1

[6] Zezhou Cheng, Qingxiong Yang, and Bin Sheng. Deep colorization. In *Proceedings of the IEEE international conference on computer vision*, pages 415–423, 2015. 1

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 3

[8] Aditya Deshpande, Jason Rock, and David Forsyth. Learning large-scale automatic image colorization. In *Proceedings of the IEEE international conference on computer vision*, pages 567–575, 2015. 3

[9] Aditya Deshpande, Jiajun Lu, Mao-Chuang Yeh, Min Jin Chong, and David Forsyth. Learning diverse image colorization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6837–6845, 2017. 2

[10] Kevin Frans. Outline colorization through tandem adversarial networks. *arXiv preprint arXiv:1704.08834*, 2017. 2

[11] Sergio Guadarrama, Ryan Dahl, David Bieber, Mohammad Norouzi, Jonathon Shlens, and Kevin Murphy. Pixcolor: Pixel recursive colorization. *arXiv preprint arXiv:1705.07208*, 2017. 2

[12] Z Hu, O Shkurat, and M Kasner. Grayscale image colorization method based on u-net network. *Int. J. Image Graph. Signal Process*, 16:70–82, 2024. 1

[13] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Let there be color! joint end-to-end learning of global and local image priors for automatic image colorization with simultaneous classification. *ACM Transactions on Graphics (ToG)*, 35(4):1–11, 2016. 2, 3

[14] Amir Kolaman and Orly Yadid-Pecht. Quaternion structural similarity: a new quality index for color images. *IEEE Transactions on Image Processing*, 21(4):1526–1536, 2011. 4

[15] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Learning representations for automatic colorization. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 577–593. Springer, 2016. 2

[16] Bin Li, Yi Lu, Wei Pang, and Huixin Xu. Image colorization using cyclegan with semantic and spatial rationality. *Multimedia Tools and Applications*, 82(14):21641–21655, 2023. 2

[17] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a “completely blind” image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012. 3

[18] Kamyar Nazeri, Eric Ng, and Mehran Ebrahimi. 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*, pages 85–94. Springer, 2018. 2

[19] Gokhan Ozbulak. Image colorization by capsule networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2019. 2

[20] Karen Panetta, Chen Gao, and Sos Agaian. Human-visual-system-inspired underwater image quality measures. *IEEE Journal of Oceanic Engineering*, 41(3):541–551, 2015. 4

[21] Georg Petschnigg, Richard Szeliski, Maneesh Agrawala, Michael Cohen, Hugues Hoppe, and Kentaro Toyama. Digital photography with flash and no-flash image pairs. *ACM transactions on graphics (TOG)*, 23(3):664–672, 2004. 1

[22] Hamza Shafiq and Bumshik Lee. Transforming color: A novel image colorization method. *Electronics*, 13(13):2511, 2024. 2

[23] Patricia Vitoria, Lara Raad, and Coloma Ballester. Chromagan: Adversarial picture colorization with semantic class distribution. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2445–2454, 2020. 2

[24] Shiqi Wang, Kede Ma, Hojatollah Yeganeh, Zhou Wang, and Weisi Lin. A patch-structure representation method for quality assessment of contrast changed images. *IEEE Signal Processing Letters*, 22(12):2387–2390, 2015. 4

[25] Primož Weingerl and Dejana Javoršek. Theory of colour harmony and its application. *Tehnički vjesnik*, 25(4):1243–1248, 2018. 4

[26] Seungjoo Yoo, Hyojin Bahng, Sunghyo Chung, Junsoo Lee, Jaehyuk Chang, and Jaegul Choo. Coloring with limited data: Few-shot colorization via memory augmented networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11283–11292, 2019. 2

[27] Tomoki Yoshida, Kazutoshi Akita, Muhammad Haris, and Norimichi Ukita. Image super-resolution using explicit perceptual loss. *arXiv preprint arXiv:2009.00382*, 2020. 3

[28] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, pages 649–666. Springer, 2016. 1, 3

[29] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 3

[30] Jiaojiao Zhao, Jungong Han, Ling Shao, and Cees GM Snoek. Pixelated semantic colorization. *International Journal of Computer Vision*, 128:818–834, 2020. 3

[31] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. *Advances in neural information processing systems*, 27, 2014. 3