From Gray to Vivid: Methods and Metrics for Image Colorization

Beaula Mahima V MA21BTECH11002
Riya Ann Easow EE21BTECH11044
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. In our work, we compare various models and strategies including curriculum learning, adding perceptual loss and ensemble methods, focusing on their performance across structural and perceptual metrics. Our results demonstrate the trade-offs between computational efficiency and output quality, with the ensemble method achieving the best performance at the cost of increased complexity.

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 relied on simple convolutional architectures with minimal skip connections, leading to slower learning and higher data requirements. Recent advancements leverage CNNs and deep learning techniques to enhance automation, semantic understanding, and performance across different image modalities.

Cheng et al. [5] introduced Deep Colorization, the first CNN-based model for automatic colorization, eliminating manual input by mapping grayscale features to chrominance values using a large-scale dataset. A joint bilateral filtering step [19] was employed to reduce artifacts. Zhang et al. [25] proposed Colorful Image Colorization, a fully automatic CNN-based approach incorporating class rebalancing to enhance rare colors and prevent desaturated outputs. It also demonstrated colorization as an effective pretext task for self-supervised feature learning, achieving state-of-the-art representation learning performance.

Carlucci et al. [4] introduced Deep Depth Colorization for depth image colorization, enhancing object recognition by learning depth-to-RGB mappings. The $(DE)^2CO$ algorithm outperformed traditional handcrafted approaches, achieving significant performance gains. Hu et al. [10] developed a U-Net-based grayscale image colorization model operating in the Lab color space, where the L component

predicts the a and b chrominance values. This method provides a foundation for further advancements in automatic colorization using deep architectures.

2.2. Diverse Colorization Networks

Deep learning based techniques for image colorization primarily employ GANs and VAEs, with methods generating either multiple diverse colorizations or a single realistic output based on semantic information.

Cao et al. [3] introduced unsupervised diverse colorization using conditional GANs, incorporating noise channels to ensure multiple plausible outputs. Nazeri et al. [17] proposed ICGAN, a fully convolutional network-inspired model that predicts a single realistic colorization using a modified generator loss for improved quality.

Frans [8] developed Tandem Adversarial Networks for line art colorization, utilizing two adversarial networks for color prediction and shading, enhanced by skip connections and L2 or adversarial losses. Deshpande et al. [7] employed a VAE-Mixture Density Network (MDN) to generate diverse colorizations, leveraging specificity, colorfulness, and gradient-based losses.

ChromaGAN [22] utilized self-supervised learning with PatchGAN to predict a single realistic colorization, integrating chrominance prediction and class distribution modeling. MemoPainter [23] incorporated memory networks with GAN-based colorization, using threshold triplet loss to enhance rare object color retrieval.

Li et al. [14] proposed SS-CycleGAN, employing dual generators and discriminators with self-attention and cascaded dilated convolutions for improved structural consistency. Shafiq and Lee [21] introduced a Transformer-GAN hybrid, where a VGG-based encoder captured global semantics, and a Swin Transformer processed color-specific features, refining outputs with perceptual, adversarial, and color losses.

These methods demonstrate diverse architectural strategies to enhance image colorization, optimizing for realism, diversity, and semantic coherence.

2.3. Multi-path Networks

Multi-path networks facilitate learning diverse feature representations but incur higher computational costs.

Iizuka et al. [11] proposed 'Let There Be Color,' a CNN-based model integrating global and local features for resolution-independent image colorization via a joint classification-colorization loss. Larsson et al. [13] utilized hypercolumns from VGG16 to predict per-pixel hue and chroma distributions, employing KL divergence loss for improved semantic coherence.

PixColor [9] introduced a two-stage approach combining a conditional PixelCNN for low-resolution chroma prediction with a refinement CNN for high-resolution colorization. ColorCapsNet [18] enhanced CapsNet with VGG-19 feature extraction, batch normalization, and capsule reduction, learning color distributions in the CIE Lab space.

Pixelated [26] employed a dual-branch network for color embedding and semantic segmentation, utilizing a conditional PixelCNN and multi-scale atrous spatial pyramid pooling. Mohammad et al. [2] proposed a tree-structured network generating multiple color hypotheses per pixel, leveraging a shared convolutional trunk for efficient color reconstruction.

These methodologies exemplify advancements in deep learning-based colorization, leveraging diverse architectures to enhance quality, efficiency, and semantic accuracy.

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 have been explored.

However, a number of image colorization models in the literature have been assessed through human evaluation, which remains the gold standard to this day. For example, in [25], 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 [6, 11].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. Project Proposal

In this project, we aim to analyze state-of-the-art (SOTA) architectures and explore ways to enhance their performance. The following are some of the ways we plan to explore to achieve our goal.

3.1. Monotonic Curriculum

Previous colorization approaches have trained models by simply mixing different datasets together. One possible approach to improve the state of the art is to incorporate a monotonic curriculum strategy to help us control the level of difficulty the model sees per epoch. We have used 2 different datasets in the following experiments: **custom ImageNet** and **COCO-Stuff**. Our key idea is based on the nature of these datasets:

- ImageNet primarily focuses on a single object per image, making it well-suited for learning object-specific colors.
- COCO-Stuff, being an object detection dataset, captures

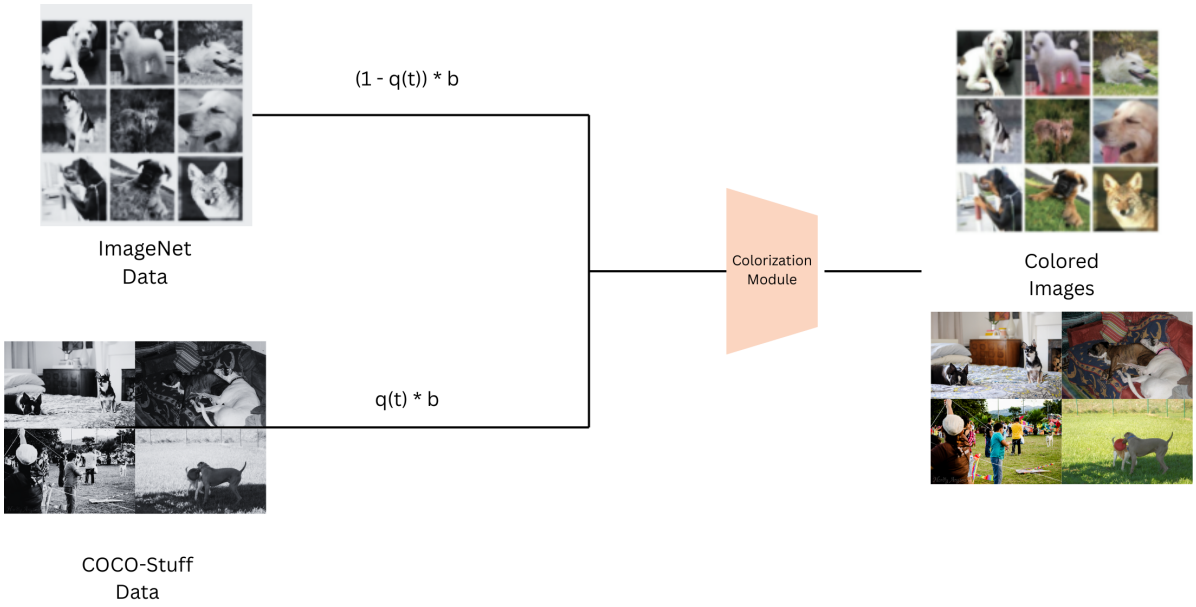

Figure 1. Pipeline for Monotonic Curriculum Strategy for Image Colorization implemented for training the Colorization model. Here, b is the batch size used during training. Finally, from the colored images, loss is computed to update the model.

more complex scenes with a broader view, better representing natural color distributions.

We have selectively chosen a subset of ImageNet so that it contains the same classes as those present in COCO-Stuff dataset. The final size of the ImageNet data is comparable to the COCO-Stuff data. Here, ImageNet data would be considered 'easy' and the COCO-Stuff data would be considered 'hard'.

We have also experimented with 2 Curriculum Learning strategies:

1. **Curriculum Learning (CL)**
   Following a curriculum learning strategy proposed in [15] in the field of deepfake detection, we perform a gradual data transition strategy that adjusts the proportion of samples from each dataset over the course of training according to

$$q(t) = \sin(t/\epsilon) \tag{1}$$

   where t is the current training epoch and $\epsilon = 2T/\pi$, ensuring a smooth and monotonic increase in COCO-Stuff data over time. Let the total number of images in a batch be b. The number of images from the ImageNet dataset will be given by $(1 - q(t)) * b$ and the number of images from the COCO-Stuff dataset will be given by $q(t) * b$. Thus, in the initial stages, the model will be trained on mostly ImageNet data. As the number of epochs progresses, images from the COCO-Stuff data will be introduced to the model. This approach helps the model learn the color distributions of the individual objects first and then learn a more general distribution of the color in images.
2. **Challenge Based Curriculum Learning (CBCL)**
   Standard CL uses a pacing function to determine the maximum level of difficulty a model should face at each step. However, this can make the model more biased towards easy data [16]. Thus, at each stage, 50% of the data is allocated to the hardest available data while the remaining 50% is allocated according to Eq. 1.

3.2. Image Aesthetic Metrics for Evaluation

Most colorization studies in the literature primarily use metrics that capture structural similarity (e.g., SSIM) or perceptual quality (e.g., PCQI, LPIPS). However, a key application of image colorization is enhancing the "aesthetic appeal" of grayscale historical images. Therefore, we believe that incorporating aesthetic quality metrics alongside structural and perceptual evaluation measures would provide a more comprehensive assessment. We plan to use the following three metrics, inspired by the ideas presented in [24], to capture the aesthetic quality of our colorizations.

3.2.1. Colorfulness Index (CI)

The Colorfulness Index (CI) measures the vibrancy and aesthetic appeal of images by analyzing their chromatic properties. It evaluates two key aspects: **chromatic contrast**, which captures variability in color differences (red-green and yellow-blue channels), and **saturation**, which reflects the intensity of colors. Higher CI values correspond to more visually striking and vibrant images, making it an effective

tool for assessing aesthetic improvements in tasks like image colorization.

The CI is calculated using the formula:

$$\text{CI} = \sqrt{\sigma_{rg}^2 + \sigma_{yb}^2} + 0.3 \times \sqrt{\mu_{rg}^2 + \mu_{yb}^2}$$

where:

- σ_{rg} and σ_{yb}: Standard deviations of red-green ($rg = |R - G|$) and yellow-blue ($yb = |0.5(R+G) - B|$) differences, representing chromatic contrast.
- μ_{rg} and μ_{yb}: Mean values of rg and yb differences, representing color saturation.
- The weighting factor 0.3 balances the contribution of saturation relative to contrast.

3.2.2. Color Harmony Metric

The Color Harmony metric is designed to evaluate the aesthetic quality of an image based on the distribution of hues in its color palette. It operates under the principle that harmonious images tend to have a more cohesive and balanced arrangement of colors, which can be quantified by analyzing the variance in their hue values. The metric converts the image to HSV (Hue, Saturation, Value) color space and calculates the variance of the hue channel. A lower hue variance indicates a more harmonious color scheme, while higher variance suggests greater diversity or imbalance in the hues.

The harmony score is mathematically defined as:

$$\text{Harmony Score} = \frac{1}{1 + \sigma_h^2} \tag{2}$$

Where σ_h^2 is the variance of the hue values in the image. The score is inversely proportional to hue variance, ensuring that images with lower hue variability receive higher scores. This metric provides a simple yet effective way to assess color harmony, with scores typically ranging between 0 and 1, where values closer to 1 indicate stronger harmony.

3.2.3. Color Distribution Balance Metric

The Color Distribution Balance metric evaluates the uniformity of color distribution in an image by calculating the entropy of the hue histogram in the HSV color space. A higher entropy indicates a more balanced and diverse color distribution.

3.3. Ensemble with a Feature Extraction Module

Various SOTA models have been developed, each excelling in different types of images (e.g. natural scenes, portraits, or historical images). However, a single model often struggles to generalize across all types of inputs. A potential solution is to employ an ensemble-based approach that combines multiple SOTA colorization models with a feature-based selection mechanism. This ensemble is designed to adaptively weight each model's output according to the characteristics of the input image, thereby enhancing overall performance and generalizability.

In our approach, we consider two colorization models for our ensemble - Colorful Image Colorisation [25] and CGAN with U-Net [12]. We employ a CLIP (ViT-B/32) encoder as the feature extractor, followed by a selector module implemented as a linear layer with a softmax activation. This module generates a set of suitability scores, indicating the relevance of each colorization model for the given input. During training, we freeze the parameters of all pre-trained colorization models and the feature extractor. Only the selector module is trained, with gradients backpropagated using a custom loss function.

The loss is defined as a weighted sum of mean squared errors (MSE) between the outputs of each colorization model and the ground truth image, where each MSE term is weighted according to the model's predicted suitability score:

$$\mathcal{L} = \sum_{i=1}^{n} w_i \cdot \text{MSE}(C_i(x), y)$$

where w_i is the selector's softmax-assigned weight for the i^{th} model, $C_i(x)$ is the colorization produced by model i for input x, and y is the ground truth color image.

During inference, the ensemble outputs the colorization result from the model with the highest predicted suitability score.

3.4. VGG-based Perceptual Loss Metric

In image colorization, relying solely on pixel-wise losses like MSE often leads to blurry or desaturated results, as these losses cannot capture high-level semantic information. To address this, we incorporated a VGG-based perceptual loss as a regularization term in the overall loss function. [20]

The perceptual loss compares deep features extracted from a pre-trained VGG network for both the predicted and ground truth color images. These features, taken from intermediate layers, capture textures and object-level details. The loss is defined as:

$$\mathcal{L}_{\text{perceptual}} = \sum_{l} \|\phi_l(I_{\text{pred}}) - \phi_l(I_{\text{gt}})\|_2^2$$

We trained a conditional Generative Adversarial Network (cGAN) for image-to-image translation in two stages. Initially, the generator was pretrained for 30 epochs using only pixel-wise supervision, minimizing the L1 loss between the generated and target images to produce structurally accurate outputs. Subsequently, the model was trained for 70 additional epochs using a combined objective that includes adversarial loss, L1 loss, and a perceptual loss based on a pretrained VGG19 network. The overall generator loss is given by:

$$\mathcal{L}_G = \mathcal{L}_{\text{GAN}} + \lambda_{\text{L1}} \cdot \mathcal{L}_{\text{L1}} + \lambda_{\text{VGG}} \cdot \mathcal{L}_{\text{VGG}}$$

where $\lambda_{\text{VGG}} = 10.0$ controls the contribution of the perceptual loss, and $\lambda_{\text{L1}} = 100.0$ controls the contribution of the direct L1 pixel loss between the ground truth image and generated image. We chose $\lambda_{\text{VGG}} = 10$ to ensure that perceptual similarity, captured via deep features from a pretrained VGG network, influences the generator meaningfully without dominating the overall loss. A higher value might overly bias the network toward matching VGG features, potentially ignoring pixel-level accuracy and introducing artifacts, while a lower value would reduce the perceptual impact, making the inclusion of VGG loss negligible. We retained $\lambda_{\text{L1}} = 100$ as used by the original authors to maintain strong structural fidelity in the reconstructions. This regularization encourages the model to produce colorizations that are not just pixel-accurate but also perceptually realistic and semantically consistent.

To compute $\mathcal{L}_{\text{VGG}}$, both the generated and ground truth images are first converted from the LAB to the RGB color space. They are then passed through a frozen VGG19 network, and the L1 distance is computed between feature activations at selected intermediate layers (e.g., `relu1_2`, `relu2_2`, `relu3_3`). This loss encourages perceptual similarity by aligning semantic content and texture. The inclusion of VGG-based loss significantly improved qualitative results and training stability.

3.5. Pretraining with L1 Loss (for GANs)

Most recent architectures in colorization literature are GAN-based, but a key challenge is the instability of GAN training, where the generator initially produces low-quality outputs and the discriminator provides weak feedback. A solution to this, inspired by SRGAN (Ledig et al., 2017), is to pretrain the generator with L1 / L2 loss before adversarial training. This allows the generator to first learn a stable grayscale-to-color mapping, preserving structural details before adversarial loss refines the realism of the generated images. This approach enhances training stability, prevents mode collapse, and improves colorization quality. Additionally, initializing the generator's encoder with a pretrained model like ResNet further strengthens performance by leveraging learned feature representations, ensuring a more robust starting point.

4. Experiments

4.1. Evaluation Metrics

Since our analysis heavily relies on the evaluation metrics chosen, we first assess their effectiveness in capturing colorization through a simple validation test. Specifically, we select two sets of images—one with high-quality colorization and another with relatively poor colorization. If the metrics yield distinguishable results between these two sets, we consider them reliable indicators of colorization.

Metric	Poor Colorization	Good Colorization
SSIM ↑	0.769	0.83
PCQI ↑	1.59	1.87
Colorfulness ↑	148.74	137.26
Color Harmony ↓	0.014	0.0089
Distribution Balance ↑	1.63	1.64
FID ↓	89.246	48.67

Table 1. Comparison of evaluation metrics between poorly and well colorized images.

Thus, we conclude the evaluation metrics we have chosen are reliable when viewed together.

4.2. Comparison of Baseline Models

The first step towards progressing with the above proposed approaches is to get a better understanding of the SOTA models in literature. For the same, we have tested out the following two models for the image colorization task.

1. **Model 1: Colorful Image Colorisation** A CNN-based model treating colorization as classification over 313 Lab-space bins, trained on ImageNet. Uses reweighted loss and annealed-mean decoding to enhance vibrancy and reduce color bias.
2. **Model 2: CGAN with U-Net** A U-Net-based conditional GAN with a PatchGAN discriminator, trained on COCO dataset. Uses a pretrained ResNet18 encoder and L1 pretraining to improve stability and color realism.

Table 2. Performance Comparison.

Metrics	Model 1	Model 2
SSIM ↑	0.86 ± 0.060	0.85 ± 0.072
PCQI ↑	1.44 ± 0.37	1.76 ± 0.43
Colorfulness ↑	115.61 ± 46.64	142.84 ± 33.11
Color Harmony ↓	0.004 ± 0.009	0.002 ± 0.005
Color Distribution Balance ↑	1.53 ± 0.13	1.68 ± 0.14

The evaluation of both the models was done on 10000 images each of the ImageNet and COCO datasets. For a more comprehensive assessment, we include metrics like PCQI and SSIM along with the aforementioned aesthetic metrics in our evaluation.

4.3. Comparsion with Proposed Approaches

The CL model consistently outperforms the baseline in SSIM and LPIPS across both datasets, indicating better structural and perceptual quality. While slightly less colorful, it maintains similar color harmony and balance, showing that contrastive learning improves generalization and

Table 3. Performance Comparison on ImageNet.

Metrics	Baseline	CL	CBCL	Perceptual	Ensemble
SSIM ↑	0.863 ± 0.076	0.868 ± 0.073	0.851 ± 0.077	0.861 ± 0.074	0.865 ± 0.064
Colorfulness ↑	142.618 ± 33.849	140.294 ± 34.473	144.481 ± 31.472	134.871 ± 33.763	138.715 ± 31.677
Color Harmony ↓	0.002 ± 0.005	0.002 ± 0.005	0.001 ± 0.003	0.001 ± 0.007	0.002 ± 0.007
Color Balance ↑	1.665 ± 0.148	1.667 ± 0.157	1.707 ± 0.144	1.684 ± 0.155	1.681 ± 0.146
PCQI ↑	1.736 ± 0.422	1.753 ± 0.476	1.759 ± 0.450	1.729 ± 0.435	1.795 ± 0.464
LPIPS ↓	0.214 ± 0.086	0.212 ± 0.084	0.227 ± 0.087	0.229 ± 0.087	0.243 ± 0.080
FID Score ↓	41.475	41.669	44.806	48.639	41.721

Table 4. Performance Comparison on COCO.

Metrics	Baseline	CL	CBCL	Perceptual	Ensemble
SSIM ↑	0.830 ± 0.067	0.837 ± 0.066	0.827 ± 0.067	0.838 ± 0.065	0.855 ± 0.067
Colorfulness ↑	143.067 ± 32.363	142.322 ± 33.263	144.436 ± 31.233	137.367 ± 34.907	137.096 ± 36.207
Color Harmony ↓	0.001 ± 0.004	0.002 ± 0.007	0.001 ± 0.007	0.002 ± 0.022	0.004 ± 0.028
Color Balance ↑	1.689 ± 0.141	1.690 ± 0.153	1.715 ± 0.147	1.697 ± 0.153	1.664 ± 0.175
PCQI ↑	1.792 ± 0.434	1.791 ± 0.450	1.801 ± 0.433	1.797 ± 0.455	1.874 ± 0.469
LPIPS ↓	0.243 ± 0.067	0.237 ± 0.066	0.244 ± 0.067	0.234 ± 0.063	0.187 ± 0.060
FID Score ↓	60.684	62.553	61.131	65.386	58.110

visual realism. CBCL enhances colorfulness and balance over the baseline, with marginal gains in PCQI. However, it slightly lags behind CL in SSIM and LPIPS, reflecting a trade-off: CBCL prioritizes vividness and visual contrast, while still preserving fidelity. CL excels in structural and perceptual similarity, whereas CBCL produces richer, more colorful images. The choice depends on application needs - CL is better for preserving fine details, CBCL for enhancing visual appeal.

The perceptual model improves PCQI and color balance, offering more natural outputs. However, it reduces colorfulness and slightly worsens LPIPS, making it better suited for realistic, subdued reconstructions over vibrant ones. The ensemble outperforms all models in SSIM, LPIPS, and PCQI, combining the strengths of individual models to deliver the most compelling results. It balances structure, color, and realism effectively.

While the ensemble achieves the best overall performance, it comes at a higher computational cost. In contrast, CL and CBCL are more efficient: CL for structure and fidelity, CBCL for color and richness. The choice ultimately depends on the trade-off between quality and efficiency.

4.4. Visualization

5. Conclusion and Future Work

From the experiments conducted, it is evident that each proposed approach has its own unique strengths. On average, the ensemble method delivers the best performance across most quantitative metrics, although it introduces additional complexity. Future work could explore dynamic curriculum learning strategies, where a policy model determines which ImageNet class would optimize learning at any given time. Additionally, the ensemble approach could be further enhanced by incorporating a more diverse set of models to boost its performance

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. 2

[3] 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

[4] Fabio Maria Carlucci, Paolo Russo, and Barbara Caputo. ˆ2 co: Deep depth colorization. *IEEE Robotics and Automation Letters*, 3(3):2386–2393, 2018. 1

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

[6] 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. 2

[7] Aditya Deshpande, Jiajun Lu, Mao-Chuang Yeh, Min Jin Chong, and David Forsyth. Learning diverse image

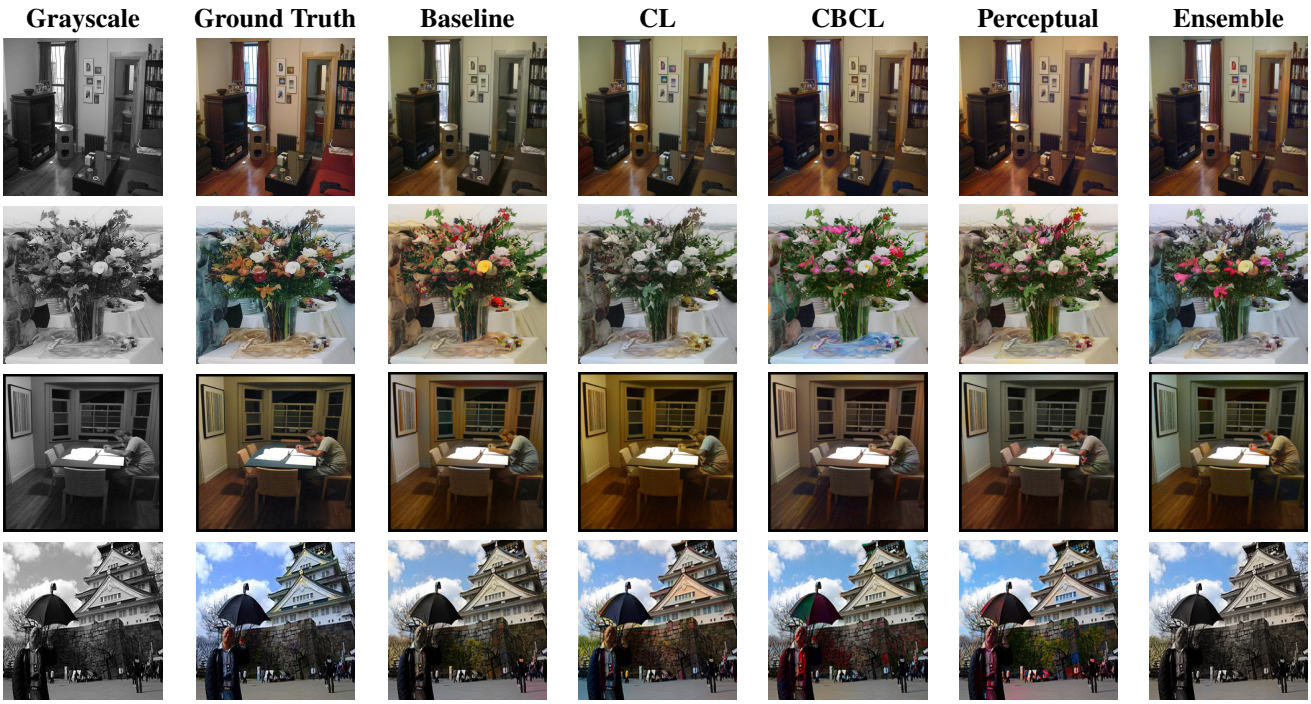

Figure 2. Comparison of grayscale colorization results across different methods for COCO and ImageNet samples.

colorization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6837–6845, 2017. 2

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

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

[10] 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

[11] 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

[12] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 4

[13] 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

[14] 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

[15] Yuzhen Lin, Wentang Song, Bin Li, Yuezun Li, Jiangqun Ni, Han Chen, and Qiushi Li. Fake it till you make it: Curricular dynamic forgery augmentations towards general deepfake detection. In *European Conference on Computer Vision*, pages 104–122. Springer, 2024. 3

[16] Yuxiao Lin, Tao Jin, Xize Cheng, Zhou Zhao, and Fei Wu. Curriculum learning aided audio-visual speech recognition with arbitrary speaker number. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2025. 3

[17] 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

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

[19] 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

[20] Rahul Sankar, Ashwin Nair, Prince Abhinav, Siva Krishna P Mothukuri, and Shashidhar G Koolagudi. Image colorization using gans and perceptual loss. In *2020 international conference on artificial intelligence and signal processing (AISP)*, pages 1–4. IEEE, 2020. 4

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

[22] 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

[23] 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

[24] Jiajing Zhang, Yongwei Miao, and Jinhui Yu. A comprehensive survey on computational aesthetic evaluation of visual art images: Metrics and challenges. *IEEE Access*, 9:77164–77187, 2021. 3

[25] 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, 2, 4

[26] Jiaojiao Zhao, Jungong Han, Ling Shao, and Cees GM Snoek. Pixelated semantic colorization. *International Journal of Computer Vision*, 128:818–834, 2020. 2