Computer-aided Colorization State-of-the-science: A Survey

Yu Cao, Xin Duan, Xiangqiao Meng, P. Y. Mok, *Member, IEEE*, Ping Li, *Member, IEEE*, and Tong-Yee Lee, *Senior Member, IEEE*

Abstract—This paper reviews published research in the field of computer-aided colorization technology. We argue that within this context, the colorization task can be considered to originate from computer graphics, advance by introducing computer vision, and progress towards the fusion of vision and graphics. Hence, we propose a specific taxonomy and organize the research work chronologically. We extend the existing reconstruction-based colorization evaluation techniques on the basis that aesthetic assessment should be introduced to ensure the computer-colored images closely satisfy human visual-related requirements. We then perform an aesthetic assessment using the proposed metric and existing evaluations, comparing the colorization performance of seven representative unconditional colorization models. Finally, we identify unresolved issues and propose fruitful areas for future research and development. Details of the project associated with this survey can be obtained at https://github.com/DanielCho-HK/Colorization.

Index Terms—Colorization, computer graphics, a taxonomy of colorization technology, colorization aesthetic assessment.

1 Introduction

Color is an integrated and crucial part of the real world. While appreciating the world's natural beauty, humans have never stopped trying to capture the rich colors of nature by utilizing methods ranging from painting to photography. A Canadian, Wilson Markle, first invented the computer-aided colorization technology in 1970 to add color to monochrome footage of the moon obtained during the Apollo program. Nowadays, such colorization-related technologies have a wide range of applications, including restoring the original colors of black-and-white photos [1], legacy films [2], cartoons [3], [4], [5], animation colorization [6], [7], [8], etc. Moreover, color has been an indispensable element of the digital world, such as in computer graphics research [9], and a significant component of computer-aided visualization of information, concepts, and ideas.

Computer-aided colorization can be defined as generating color information from gray-scale images and line drawings (or sketches) while keeping the structural details unchanged through computing technology. For more than two decades, this field has attracted the attention of many researchers in computer graphics and computer vision, who were trying to tackle four main problems: (1) Difficulty in recovering the original color information; (2) Integration of semantic understanding and different color sources; (3) Non-photorealistic colorization and (4) Evaluation of colorization.

Researchers have proposed various solutions to address these four challenges: (1) To tackle the multi-modal problem of image color [10], such as varying the colors of leaves in different seasons (from green to yellow), two leading approaches have been used: user-guided *conditional* colorization and deep color priors-based automatic *unconditional* colorization. The former provides precise color information through various forms of user interaction (e.g., reference images, color hints, palettes, and text prompts), emphasizing the user's dominant role in the coloring process, while the latter relies on prior color information learned from large-scale datasets, offering more prosperous and more accurate data modeling and generation capabilities. (2) Semantic correspondence has different implications in both conditional and unconditional coloring processes. In conditional coloring, this is primarily reflected in techniques based on reference images and user hints/scribbles. The coloring result must ensure that the color in the current area matches the color of the same semantic area in the reference image or aligns with the user-provided color information. It is also crucial to manage different sources of color during the coloring process [11], such as resolving conflicts between user-provided color information and learned color priors. The unconditional method involves image semantic understanding at three levels: global, pixel, and instance, which are analyzed in detail in the following technical review. (3) In the colorization literature, an increasing number of articles address the colorization of manga, line drawings, and sketches, in addition to gray-scale images. Chen et al. [12] provided a detailed comparison of these non-photorealistic colorization targets, which we generally refer to as line drawings in our survey. Unlike gray-scale images, line art presents more challenges, including sparse information, a lack of high-quality paired data for training, and complex line structures with unique tones and textures in anime

- *Yu Cao is with the School of Fashion and Textiles, The Hong Kong Polytechnic University, Hong Kong. E-mail: yu-daniel.cao@connect.polyu.hk.*
- *Xin Duan, Xiangqiao Meng, and Ping Li are with the Department of Computing and the School of Design, The Hong Kong Polytechnic University, Hong Kong. E-mail: {hizuka.duan, xiangqiao.meng}@connect.polyu.hk, p.li@polyu.edu.hk.*
- *P. Y. Mok is with the School of Fashion and Textiles as well as the Research Centre of Textiles for Future Fashion, The Hong Kong Polytechnic University, Hong Kong. E-mail: tracy.mok@polyu.edu.hk.*
- *Tong-Yee Lee is with the Department of Computer Science and Information Engineering, National Cheng-Kung University, Tainan 70101, Taiwan. E-mail: tonylee@ncku.edu.tw.*

Yu Cao, Xin Duan contributed equally to this work.
P. Y. Mok, Ping Li, and Tong-Yee Lee are corresponding authors.

and manga. Therefore, methods designed explicitly for line drawing colorization often utilize user inputs or reference images as a pre-condition. (4) Evaluating colorization results is challenging because they are inherently ill-posed and subjective. Ground truth is not always available. Even when gray-scale versions of a color image are constructed to create ground truth, the results can be diverse and subjective and may not necessarily align with the ground truth. Evaluation metrics dedicated to colorization need to be developed for the benefit of the research community.

Although some excellent reviews have been published on colorization, they generally have certain limitations. For example, Anwar et al. [13] mainly focused on single-image colorization, while Huang et al. [14] focused on deep learning-based colorization and largely ignored traditional methods. Chen et al. [12], on the other hand, reviewed the field of colorization, from image analogy to learning-based methods, but did not discuss the latest advancements, and their proposed taxonomy was developed according to various types of colorization objects. Comparatively, our survey addresses some of the above-mentioned weaknesses and gaps, providing a much more detailed, comprehensive, and up-to-date review of the field of computer-aided colorization. We also present our taxonomy to organize existing related work from the perspective of methodological approaches. Considering colorization as a kind of computer graphics task, we divide colorization into three broad categories: conditional methods, unconditional methods, and video colorization (see Fig. 1). This survey complements previous reviews, analyses the development of colorization technologies, and identifies potentially rewarding future directions in colorization research. Moreover, reviewing the existing work reveals that evaluation metrics for reconstruction-based models were not initially designed for the colorization task, and there is no ground truth image. The effectiveness of computerized colorization, precisely the coloring quality, is a relatively abstract concept that is hard to quantify and involves aesthetic assessment. This paper proposes a novel colorization aesthetic assessment method inspired by CLIP-based image quality assessment research. The method simulates a human vision perception system.

The main contributions of our paper can be summarized as follows:

- A thorough review of the research and other materials on colorization technology published during the past two decades provides a precise, insightful literature-based analysis for follow-up research.
- The first introduction of aesthetic assessment of colorized images and the first evaluation of seven unconditional image colorization methods.
- A discussion of the main challenges, development trends, and suggestions concerning potentially fruitful future research directions and technological advancements.

The remainder of this survey is structured as follows: Section 2 reviews the published work on colorization technology and is organized according to our proposed taxonomy. Section 3 summarizes representative datasets that can be used for training learning-based colorization models. Section 4 introduces the concept of colorization aesthetic assessment and compares seven unconditional colorization techniques based on the proposed new aesthetic assessment. Section 5 discusses future research directions, and Section 6 contains a summary and conclusions.

2 Colorization

2.1 Conditional Colorization

Conditional colorization refers to the colorization technology that can explicitly generate diverse colored results according to different types of user inputs, namely conditional controls. Conditional methods involve five main types of control, namely reference image, hint/scribble, palette, text, and multi-modal controls. The earliest natural image colorization publications are based on reference images [15] and user hints [16]. With the development in technology, the colorization methods have also evolved from the traditional non-parametric optimization paradigm to the learning-based paradigm, with the coloring objects extending from gray-scale images to line drawings, including manga [17], [18], anime [19], [20], [21], [22], cartoons [23], [24], [25], icons [26], [27], etc.

2.1.1 Reference-based Methods

Reference-based colorization methods convert color information from the reference images to the target gray-scale images or black-and-white line drawings (see Fig. 2).

Gray-scale Image Colorization. Reference-based gray-scale image colorization originates from the concept of image analogy [28], a method of automatically learning an image filter from training data. As in the case of other classic computer graphic tasks, such as texture synthesis, texture transfer, and artistic rendering, reference-based image colorization can be considered an image filter simulation based on image analogy, depending on establishing semantic correspondence between the reference and target images, i.e., identifying and aligning similar semantic features or objects in both images so that the color information from the reference image can be accurately transferred to the gray-scale image.

Optimization-based methods. Inspired by this, Welsh et al. [15] proposed the first method of transferring color from a source image to a target gray-scale image. Their basic idea was first to perform pixel neighborhood matching in the luminance channel and then to transfer chromaticity values from the source to the target. Irony et al. [29] introduced an exemplar-based colorization technique, incorporating a specially designed texture-based classifier for more accurate localized color transfer. This classifier is derived from a detailed analysis of low-level features in the reference image. In the early studies, gray-scale image colorization models were built upon the assumption that similarities in gray-scale intensities indicate color similarities. Colorization algorithms can be misled by intensity disparities arising from variations in shades and brightness between its reference image and the target image. To address the problem of illumination inconsistencies between target and reference images, Liu et al. [30] introduced an intrinsic colorization approach, which involved computing an illumination-independent reference image by means of intrinsic image decomposition. To mitigate the problems

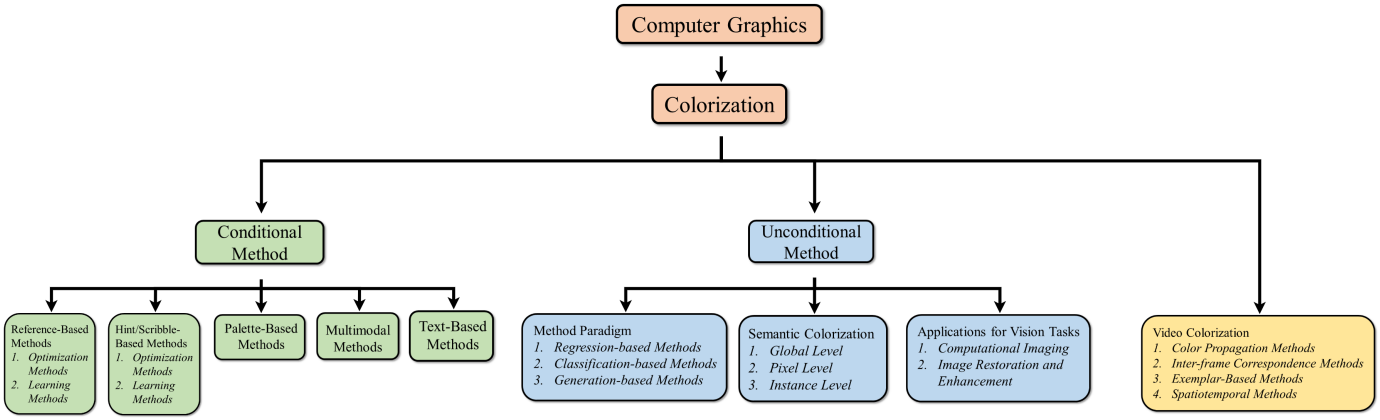

Fig. 1. Taxonomy of colorization technology. We classify colorization as a sub-task of computer graphics, in which the green sections belong to the explicit diverse colorization, using conditional controls, and the blue sections belong to the unconditional colorization methods and applications for vision tasks. Video colorization is separately shown in yellow to indicate it as an extension of image colorization in spatiotemporal dimensions.

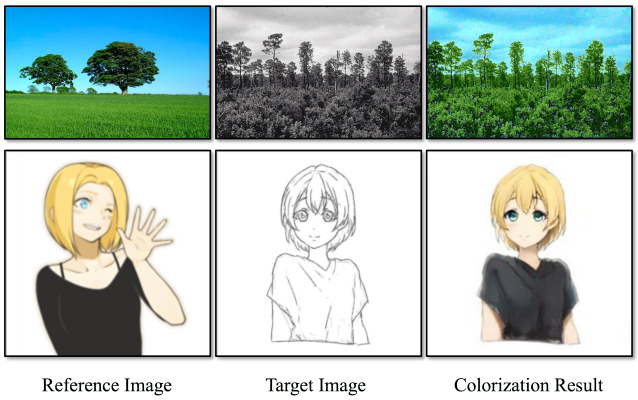

Fig. 2. Illustration of reference-based colorization. The top row shows reference-based gray-scale image colorization and the bottom row shows reference-based line drawing colorization. Images courtesy of [15], [19].

caused by variations in illumination, their method leveraged multiple reference images obtained from the internet. A similar concept of using an extensive online image database to colorize gray-scale images was also proposed by Morimoto et al. [31]. The success of such reference-based methods, however, depends heavily on retrieving a suitable reference image. To expand the colorization target range from famous landmarks to general objects and scenes, Chia et al. [32] proposed a more user-friendly colorization method that utilized semantic labels to search appropriate reference images from the internet and thereby facilitate accurate color transfer. To improve the speed of image colorization, Gupta et al. [33] proposed a method that used a fast cascade feature matching scheme for rapidly finding correspondence between reference images and target images at the superpixel level. Bugeau et al. [34] designed a variational model simultaneously modeling color selection and spatial constraints. Li et al. [35] presented a novel approach to colorizing target superpixels by formulating it as a problem of dictionary-based sparse reconstruction. They introduced the first sparse pursuit image colorization method, using a single reference image. By reducing limitations in terms of the images provided by users and improving the precision of dense feature matching between images, it can make reference-based colorization technology more versatile. It generally presents a challenge when users provide reference and target images of objects with different scales. To tackle the problem of feature matching involving different scales, Li et al. [36] introduced a cross-scale matching technique that incorporates localized considerations of various potential scales during the matching process. They subsequently employed the graph-cut technique to fuse the scales globally, aiming to identify spatially coherent scales that exhibit high-quality matching. Fang et al. [37] approached exemplar-based image colorization from the point of view that it was a problem of color selection, incorporating regularization constraints. They focused on utilizing superpixels as processing units so as to improve both the efficiency and robustness of colorization. Notably, their work introduced the utilization of superpixel-based non-localized self-similarity and localized spatial consistency as novel techniques for image colorization.

In summary, the technological developmental route of reference-based gray-scale image colorization was first started with single image analogy, then utilizing the retrieval-based method for selecting appropriate reference images when involving multiple reference images, until the superpixel became the most effective solution among the traditional techniques.

Learning-based methods. Since 2012, with the successful application of Deep Convolutional Neural Network (CNN) models [38] for many vision-related studies, reference-based image colorization methods have also changed from traditional optimization-based non-parametric models to CNN-based parametric models. To address more complex situations where clear correspondence between the target and reference images is lacking, He et al. [39] proposed an end-to-end network to learn the selection, propagation, and prediction of colors from existing data. Unlike earlier deep learning-based colorization approaches, their method effectively captured local and global contents, resulting in improved colorization outcomes, which avoided semantic inaccuracies or low color saturation. Xiao et al. [40] presented a Dense Encoding Pyramids Network for colorization, which mapped the color distribution from a reference image onto a gray-scale image. Their model incorporated

novel Parallel Residual Dense Blocks to capture comprehensive local-global context information. Furthermore, a Hierarchical Decoder-Encoder Filter was employed to aggregate color distribution information across adjacent feature maps. Inspired by work of stylizing photorealistic images [41], [42], Xu et al. [43] designed a deep learning model consisting of a transfer net and a colorization net to perform real-time exemplar-based image colorization. For a reliable reference-based image colorization system, the semantic colors associated with the objects and global color distributions are important color characteristics of the reference image. Lu et al. [44] introduced Gray2ColorNet to address this aspect. This end-to-end deep neural network innovatively combined reference image semantics and global colors for effective image colorization. Yin et al. [11] modeled colorization as a query-assignment problem for different color sources and designed a unified attention mechanism framework. Under this unified framework, selecting and assigning colors from a reference image to the gray-scale image adhered to a shared criterion, utilizing the semantic features as a primary factor. Inspired by the observation that a broad learning system was capable of efficiently extracting semantic features [45], Li et al. [46] proposed Broad-GAN as an approach for semantic-aware image colorization. They devised a customized loss function to improve training stability and to evaluate the semantic similarity between the target and ground-truth images. Attention-aware methods have recently emerged to tackle the inherent semantic correspondence problem in reference-based image colorization. Carrillo et al. [47] introduced a super-attention block that leveraged superpixel features to transfer semantically related color characteristics from a reference image across various scales within a deep learning network. Similarly, Bai et al. [48] proposed a Semantic-Sparse Colorization Network, which employed a sparse attention mechanism to transfer global image styles and detailed semantic-related colors to gray-scale images in a coarse-to-fine manner. Wang et al. [49] developed an effective exemplar-based colorization strategy utilizing a pyramid dual non-local attention network to explore long-range dependencies and multi-scale correlations.

In summary, Learning-based reference colorization research mainly tackled the semantic correspondence problem by introducing different solutions, including local-global deep features aggregation, exemplar-based style transfer formulation, and attention mechanism modules. In addition, the decoupling of the color sources when performing the reference-based colorization is of significance. Usually, the color comes from 1) the semantic colors linked to objects in the reference image, 2) the global color distribution encompassing tones and hues of the reference image, and 3) the color information learned from large-scale datasets.

Line Drawing Colorization. The early method [50] formulated reference-based line drawing colorization as a neural style transfer problem [51]. Since basic neural style transfer models designed for natural images cannot deal with line drawings, they incorporated a residual U-Net architecture and utilized an Auxiliary Classifier Generative Adversarial Network (AC-GAN) [52] to apply the desired style to the gray-scale sketch. In line drawing colorization, a specialized group of researchers focuses on 'manga' and 'anime,' namely comics and animations originating in Japan. Furusawa et al. [18] introduced Comicolorization as the pioneering approach to colorizing complete manga titles, encompassing sets of manga pages. The semi-automatic system uses reference images to colorize the input black-and-right manga images. In the anime creation industry, artists often manually draw anime character illustrations with empty pupils first and then only thereafter fill in the preferred colors or details in the pupils. Akita et al. [53] introduced a colorization model, combined with a pupil position estimation module, to colorize anime character faces automatically with accurate pupil colors. To solve the problem of not being able to obtain paired training data before and after colorization, Lee et al. [54] proposed a training scheme aimed at learning visual correspondence. They achieved this by generating self-augmented references with a self-supervised training scheme, eliminating the need for manually annotated labels of visual correspondence. This development facilitated end-to-end network optimization without explicit supervision. Cao et al. [55] developed a segmentation fusion model to reduce the color-bleeding artifacts effectively. Li et al. [56] identified the problem gradient conflict within the attention modules during line-art colorization, negatively impacting the training stability. To address this issue, they proposed a training strategy called Stop-Gradient Attention. This strategy eliminated the gradient conflict problem, enabling the model to learn improved colorization correspondence. Liu et al. [24] employed a multi-scale discriminator to enhance the visual realism of colorized cartoons, focusing on improving both global color composition and local color shading. Chen et al. [25] introduced an active learning-based framework that combined the local-region matching of line art and reference-colored images, followed by spatial context refinement using mixed-integer quadratic programming (MIQP). Cao et al. [19] devised an explicit attention-aware model for generating high-quality colored anime line drawing images. Wu et al. [57] proposed a pioneering self-driven dual-path framework for reference-based line art colorization based on limited data. More recently, Cao et al. [21] introduced AnimeDiffusion, the first diffusion model tailored explicitly for reference-based colorization of anime face line drawings.

In summary, reference-based line drawing colorization research focused on cross-domain semantic correspondence and color consistency to generate appropriate colored results. Nevertheless, due to the lack of pairs of high-quality training data, it is difficult to train models in a supervised manner. Since there is a lack of large-scale training data, like natural images, the generalization of existing models is still limited.

2.1.2 Hint/Scribble-based Methods

Hint/Scribble-based colorization is a technology that performs colorization by propagating local user-provided color scribbles (Fig. 3) and color points (Fig. 4).

Gray-scale Image Colorization. Levin et al. [16] introduced an innovative interactive colorization technique that assumed neighboring pixels in space-time, with similar intensities, should exhibit similar colors. Using a quadratic cost function, they formulated the color propagation process as an optimization problem. Manual scribbling can be tedious and time-consuming for images with complex details and

requires aesthetic-related skills to obtain realistic results. The method developed by Irony et al. [29] can automatically generate 'micro-scribbles' from the image the user provides as an example, greatly facilitating user involvement. To solve the 'color bleeding' problem in boundary regions, Huang et al. [58] developed an adaptive edge detection scheme to prevent colorization from bleeding over boundaries. Yatziv and Sapiro [59] designed a fast colorization framework based on the concept of color blending to speed up the colorization process, and their method can be easily extended to video colorization without the need for optical flow computation. Since previous user-guided methods require a vast number of user inputs (e.g., in the form of strokes) to achieve high-quality colorization of images with complex textures, Luan et al. [60] considered the colorization problem as one of image segmentation by using the concept of texture cluster, and proposed an easy two-step colorization system, including Color Labeling and Color Mapping.

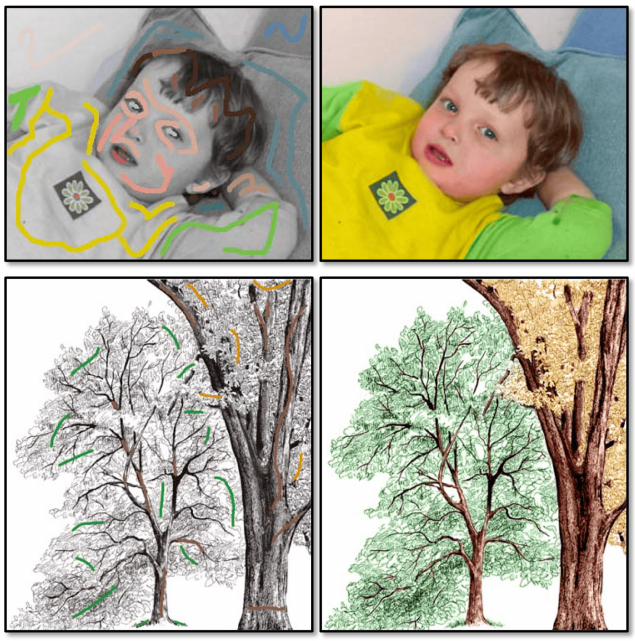

Fig. 3. Illustration of colorization using color scribbles (or strokes). The top row shows gray-scale image colorization, and the bottom row shows line drawing colorization. Images courtesy of [16], [17].

After 2017, user-guided colorization methods have entered the era of deep learning. Zhang et al. [62] introduced deep learning technology to assist users in making informed color input decisions. By training a model on large-scale natural image datasets, the model acquires fundamental capabilities for image semantic recognition and color information statistics. Assisted with an interactive system, users can specify their preferences using the trained model and thereby generate plausible colorization results [62]. Kim et al. [63] designed a simple add-on edge-enhancing network to enable users to interactively annotate the color bleeding region with scribbles. Instead of performing a direct operation at the image level, their model used scribbles and intermediate feature maps to generate edge-enhanced colorization outputs. Xiao et al. [64] designed a two-stage deep colorization model that simultaneously accommodated inputs of both local color points and global palette. Yun et al. [65] proposed the first color point-based real-time colorization model based on the Vision Transformer [66] which, by leveraging the self-attention mechanism, can avoid producing partially colorized outputs.

Fig. 4. Illustration of colorization using color points. The top row shows gray-scale image colorization, and the bottom row shows line drawing colorization. Images courtesy of [50], [61].

Line Drawing Colorization. User-guided colorization methods for line drawings can be traced back to the work of Qu et al. [17], who developed a stroke-based approach that was tailored specifically for colorizing manga, characterized by intensive strokes, hatching, halftoning, and screening. Due to the sparse continuity of patterns in manga, conventional intensity-continuity-based techniques are ineffective. Qu et al. [17] proposed a method to propagate the color in pattern-continuous and intensity-continuous modes automatically. Sýkora et al. [67] presented a flexible colorization tool that could be applied to various drawing styles as opposed to the previous style-limited approaches.

As in the case of natural images, starting in 2017, many user-guided line art colorization methods using deep learning techniques were developed. Sangkloy et al. [68] formulated the colorization process as a sketch-to-image synthesis problem involving scribbles control. They were the first to utilize GAN to generate realistic images according to sketches and sparse color scribbles. Liu et al. [23] designed the auto-painter model, based on conditional Generative Adversarial Networks (cGAN) [69], which automatically generates colorized cartoon images from a sketch. Ci et al. [70] proposed a model based on cGAN architecture for scribble-based anime line art colorization. To generate authentic illustrations with accurate shading, they integrated their framework with WGAN-GP [71] and perceptual loss [72]. Zhang et al. [61] proposed a learning-based framework consisting of two distinct stages: drafting and refinement. This decomposition simplified the learning process and enhanced the overall quality of the final colorization outcomes. In line art colorization, flat filling is a process that uses relatively flat colors rather than color textures. Zhang et al. [73] proposed the Split Filling Mechanism framework to control the areas defined by user scribbles and thereby generate realistic color combinations. Their method can fill in flat, consistent colors to regions instead of pixel-level color textures. Yuan and Simo-Serra [74] presented

a Concatenation and Spatial Attention module that can generate consistent and high-quality line art colorization from user inputs. Previous methods performed colorization in RGB color space, which resulted in dull colors and inappropriate saturations. To address this problem, Dou et al. [75] introduced the DCSGAN model, which was the first to utilize Hue, Saturation, and Value (HSV) color space to enhance anime sketch colorization. The HSV color space closely aligns with the human visual cognition system. It is well-suited for colorization tasks incorporating prior human drawings, including hue variation, saturation contrast, and gray contrast. Recently, Carrillo et al. [76] introduced an interactive method for colorizing line art using conditional Diffusion Probabilistic Models. Cho et al. [77] developed the GuidingPainter model to improve the efficiency of the interactive sketch colorization process, based on the concept that making the model actively seeks regions where color hints would be provided rather than rely too heavily on deciding color local-position information provided by users.

There are two primary research motivations for hint/scribble-based methods. First, they enhance the accuracy and naturalness of image coloring. User-provided color cues can serve as prior information, helping the algorithm more accurately infer the color distribution across different image areas. Second, these methods increase user engagement and satisfaction. Allowing users to provide color prompts gives them greater control, enabling them to customize the image's colors according to their preferences and needs.

In addition, it is important to note that user-hint-based colorization methods are more commonly used in non-photorealistic sketch coloring tasks. This is because, unlike natural images, line art images lack a large dataset that defines prior color distributions. Furthermore, in animation or artistic creation, many color schemes do not exist in nature. Due to the unique nature of non-photorealistic art coloring, interactive user-hint-based methods provide users with greater control and creative freedom.

Finally, to effectively manage conflicts between user-provided hints and the color priors derived from the algorithm's semantic understanding of the image, colorization systems should be designed with a flexible architecture that prioritizes user input when necessary. Additionally, incorporating feedback mechanisms that show users how their hints are being applied can help achieve the desired results while maintaining a balance between algorithmic suggestions and user creativity.

2.1.3 Palette-based Methods

Palette-based image colorization is a technique where a limited set of colors, known as a palette, is used to colorize a gray-scale image, as illustrated in Fig. 5. The color palette usually reflects the overall tones or themes of the image, while the image color conveys emotions by means of color themes. Wang et al. [78] proposed the first system, called the Affective Colorization System, which can efficiently colorize a gray-scale image semantically, using a color palette with emotional information incorporated. Bahng et al. [79] regarded the palette as an intermediate representation that conveys the semantics of the image. They first designed a model to generate multiple palettes according to different text inputs and then performed palette-based gray-scale image colorization. Xiao et al. [40] proposed a reference-based colorization framework, with the palette being one

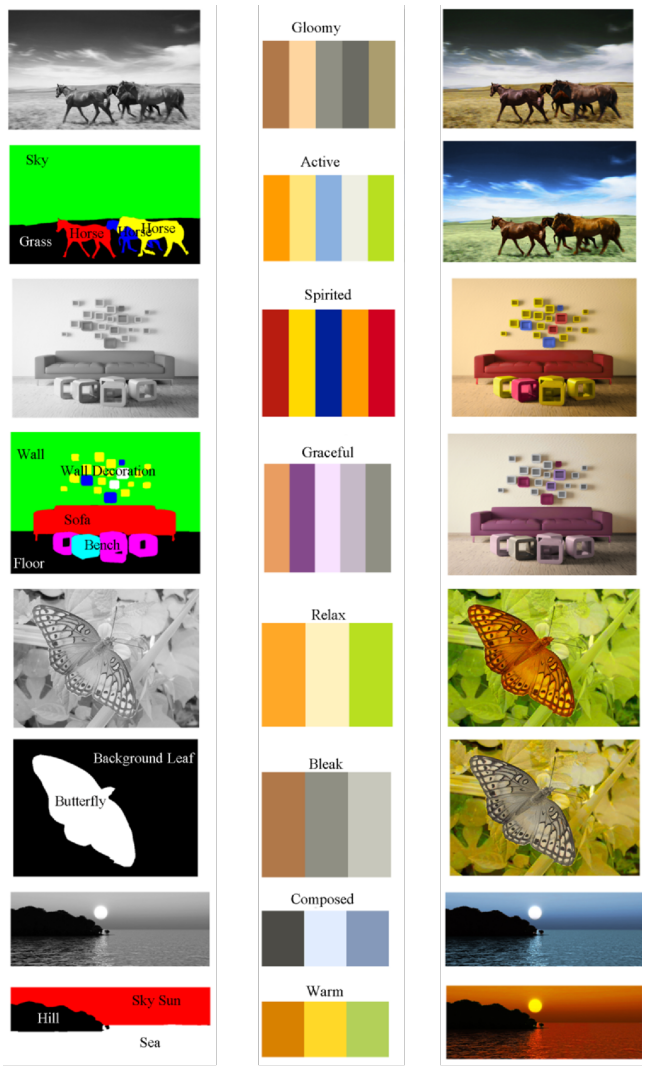

Fig. 5. Palette-based colorization. Images courtesy of [78].

reference type, to generate realistic colored outcomes. Wu et al. [80] introduced a flexible icon colorization model based on user-guided images and palettes.

Compared to other conditional colorization technologies, palette-based methods have attracted relatively less attention from researchers, with publications dealing with palette-based methods mainly covering natural gray-scale image colorization or icon sketch colorization.

2.1.4 Text-Based Methods

Text-based colorization [81] performs coloring according to users' instructions provided in the form of texts (natural languages), and being a relatively novel cross-modal interactive method, to both natural and line images, as illustrated in Fig. 6.

This technology can be categorized as non-diffusion-based and diffusion-based methods. For traditional non-diffusion-based methods. Chen et al. [82] pioneered a text-based method for colorization. They developed a comprehensive modeling framework that completes two interlinked tasks of image segmentation and colorization for language-based image editing. Their approach involved us-

ing a CNN to process the source image and an LSTM [83] network to encode the textual features according to the language-based user inputs. Another important approach was that of Bahng et al. [79], who focused on linking specific words to particular colors and thereby encapsulating the semantics of the text input. Their method produced relevant color palettes that captured the essence of the text and which were then applied to add color to gray-scale input images. They mentioned that their technique allows individuals without artistic expertise to create color palettes that effectively communicate high-level concepts. Weng et al. [86]

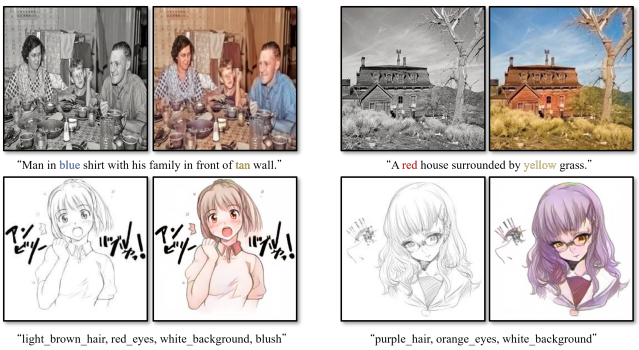

Fig. 6. Illustrations of text-based colorization. The top row illustrates gray-scale image colorization, and the bottom row line drawing colorization. Images courtesy of [84], [85].

presented a method that decoupled colors and objects into different spaces. Their approach allowed for correctly applying designated (potentially unusual) color words to objects, successfully addressing the common problems associated with coupled color and objects, and mismatches between them. Chang et al. [84] introduced the first transformer-based text-based colorization system by analyzing the cross-modal relationship between images and texts using cross-attention learning, thereby addressing the large gap between the two modalities. Early methods independently leveraged two distinct architectures for feature extraction: CNN for images and LSTM for relevant textual descriptions. In another study, Chang et al. [87] proposed a Transformer-based framework to automatically consolidate related image patches and attain instance awareness without additional information. Their method solved the problem encountered by previous researchers in differentiating between different instances of the same object.

Due to the cross-attention mechanism module, Stable Diffusion-based approaches can be easily combined with text control for image coloring tasks. Chang et al. [88] utilized the robust language understanding and extensive color priors provided by Stable Diffusion [89] for text-based colorization. Unlike previous approaches, which relied on comprehensive color descriptions for many objects in the image, this method avoided suboptimal performance, especially for items without color-matched descriptive words. Thanks to the excellent data distribution modeling performance of diffusion models and the color priors learned in large-scale image datasets, [90], [91] can perform text-based coloring tasks well and can even be extended to support other interactive methods or unconditional automatic coloring.

Unlike natural gray-scale images, which can combine color and semantic descriptions for accurate colorization, cartoon line art coloring benefits from the rich tags provided by the Danbooru dataset. These independent tags can be combined to produce coloring results with different effects. Kim et al. [85] proposed an alternative approach, employing a GAN for colorizing line art. Their method used monochromatic line drawings and color tag data as inputs to generate high-quality colored images. The current popular ControlNet-based anime content generation method [92] also utilized these tag-based prompts to perform sketch-guided anime content generation.

Text-based image colorization is a technique that involves adding color to gray-scale images or sketches using textual descriptions as guidance. It is worth noting that the forms of the two types of text guidance are different. For natural images, the text descriptions typically combine color and semantic information, whereas, for cartoon line art, tag-based prompts are mainly used to depict characters or scenes.

2.1.5 Multi-modal Methods

Multi-modal colorization methods perform colorization by combining different types of control. Huang et al. [93] presented UniColor, the first unified framework that enabled colorization in multiple modalities, encompassing both unconditional and conditional approaches and accommodating various conditions, including stroke, exemplar, text, and combinations. The framework involved a two-stage colorization process that integrated these conditions into a single model. In the initial stage, the diverse multi-modal conditions were transformed into a shared representation, known as hint points. Notably, they introduced a CLIP-based [94] approach to convert textual inputs into hint points, ensuring compatibility with other modalities. The subsequent stage involved a Transformer-based network consisting of Chroma-VQGAN and Hybrid-Transformer components. Recently, two diffusion-based models for natural image colorization [95] and anime sketch colorization [20], [96] have been proposed, both supporting multiple types of control for interactive colorization.

In summary, with the great improvement in the performance of generative models, multi-modal interactive coloring technology has gradually become the research direction and has high application potential.

2.2 Unconditional Colorization

Unconditional image colorization is a process where a gray-scale image is converted into a color image without any additional input or guidance from the user. In previous studies, researchers have also referred to this as automated colorization. In this section, we discuss the technology from three perspectives: the method paradigm, automatic semantic colorization, and applications for vision tasks.

2.2.1 Method Paradigm

Regression-based Methods. Early attempts [97], [98] leveraged neural networks, as **regression models**, to minimize average reconstruction errors, with the colorization entering the era of deep learning by using a neural model to learn the

mapping between gray-scale images and colored images. The loss function can be defined as follows:

$$\mathcal{L}_{\text{regression}} = \frac{1}{N}\sum_{i=1}^{N}\|\mathbf{y}_i - \hat{\mathbf{y}}_i\|^2, \quad (1)$$

where N is the sum of pixel numbers, $\mathbf{y}_i$ is the real color value of the ith pixel, $\hat{\mathbf{y}}_i$ is the predicted color value.

Classification-based Methods. By introducing the concept of representation learning [99], Zhang et al. [100] treated colorization as a **classification task** using the image's L channel as input and its ab channels as outputs for supervised learning. Changing pixel regression to pixel classification significantly improved the colorization quality, especially regarding saturation. The loss function can be defined as follows:

$$\mathcal{L}_{\text{classification}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}\log(\hat{y}_{i,c}), \quad (2)$$

Where C is the total number of color categories, $y_{i,c}$ is the true label of the ith pixel on the category c (usually one-hot coding), $\hat{y}_{i,c}$ is the probability that the model predicts that the ith pixel belongs to class c. Similarly, Larsson et al. [101], [102] explored colorization as a means of self-supervised learning for visual comprehension, free from the constraints of pre-trained backbone models. These insights into representation learning through colorization have significantly advanced the understanding of self-supervised tasks and their applications in computer vision.

Generation-based Methods. Different from conditional colorization methods, involving various types of user control to generate diverse colored results explicitly, automatic methods, utilizing generative models, can perform diverse coloration implicitly, as illustrated in Fig. 7. Charpiat et al. [10] undertook a pioneering work of implicitly learning the multi-modal probability distribution of colors. It is worth noting that the word 'multi-modality' used in their paper about fifteen years ago meant 'diversity of colors,' which differs from the meaning of multi-modality today, namely, multiple forms of sensory inputs, such as images, text, audio, and video. Nevertheless, as a result of their groundbreaking work [10], many subsequent studies dealing with diverse colorization still used 'multi-modal' within the context of 'diversity of colors.'

We find that previous studies, aimed at diverse colorization and published in different periods, adopted the contemporary famous generative model algorithms as the backbone of their approaches, including VAE [104], [105], Flow [106], Auto-Regressive [107], [108], [109], [110], [111], [112], [113], GAN [103], [114], [115], [116], [117], [118], [119], [120], [121], [122], [123], and Diffusion Models [90], [91], [124].

In summary, Auto-Regressive (including Transformer) and GAN models are the mainstream generation architectures for designing diverse coloring methods. Since generative algorithms aim to model the distribution of training data, they have the inherent property of generating diverse colorization results. As the performance of the diffusion model stands out in various vision tasks, its distribution-modeling performance in terms of data diversity is superior to previous generative algorithms. Hence, researchers tend to design multi-modal interaction-based coloring methods based on diffusion models.

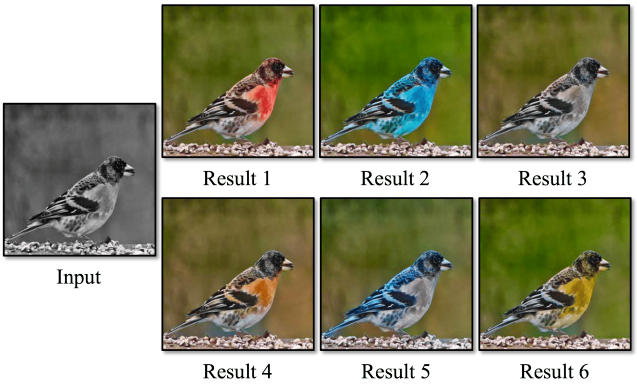

Fig. 7. Illustration of diverse colorization methods. The model, trained on a large-scale image dataset, automatically learns the color priors and performs implicit diverse colorizations. The gray-scale input image is on the left, and the six images on the right are the colored results of certain unconditional colorization methods. Images courtesy of [103].

2.2.2 Semantic Colorization

Semantic information is essential for performing high-quality colorization. We divide the existing semantic automatic colorization methods into three categories, including global level, pixel level, and instance level, which correspond to three vision tasks: image classification, semantic segmentation, and instance segmentation, as illustrated in Fig. 8.

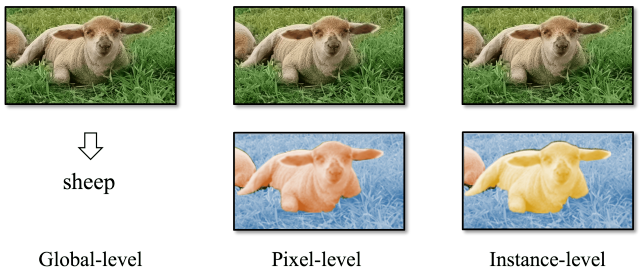

Fig. 8. Illustration of three kinds of automatic semantic colorization methods. Global level, pixel level, and instance level separately correspond to three vision tasks: image classification, semantic segmentation, and instance segmentation.

Global Level. Iizuka et al. [125] was the first to propose using semantic information for a colorization process; their approach involved a fully automated data-driven method to classify and colorize images utilizing a labeled dataset. It integrates local image features, calculated from local patches, with visual context derived from semantic class labels. This enables the model to determine and understand the correlation between semantic labels and their respective colors. Özbulak [126] proposed a colorization approach that utilized the image recognition and understanding capabilities of Capsule Networks (CapsNets) [127]. Vitoria et al. [128] introduced an approach that combines adversarial learning with semantic information for colorization. They leveraged the power of GAN to generate realistic and visually pleasing colored images. Their method incorporated semantic class distribution learning, guiding the colorization process based

on the semantic context of the image. The method proposed by Jin et al. [129] introduced a new automatic colorization approach based on a broad learning system [45], which accurately determined the category and color of pixel blocks through training. Kang et al. [113] proposed a semantically reasonable and visually vivid colorization approach that utilized a pixel decoder and a query-based color decoder. The former was responsible for reconstructing the spatial information of the image, while the latter used rich visual cues to enhance color queries without requiring manually determined color priors.

Pixel Level. The task of image classification typically prioritizes transform invariance, meaning the classification should remain the same even if the image is transformed. Nevertheless, satisfactory colorization requires the use of transform-variant features. In this regard, semantic segmentation is viewed as a more appropriate sub-task for colorization, which similarly relies on pixel-level and transform-variant representations. Sharing the same idea, Zhao et al. [130] proposed a method incorporating detailed object semantics at the pixel level to direct the process of image colorization. They trained a network with two loss functions, one for semantic segmentation and the other for colorization. By leveraging the pixel-level object semantics, their approach enhanced the precision and uniformity of the colorization process. Building on this work, Zhao et al. [131] further investigated how to employ pixel-related semantics to produce realistic colorization. In addition to segmenting objects with common illumination conditions, Duan et al. [132] proposed a method for coloring shadowed images, which can be regarded as a specialized form of segmentation that distinguishes between shaded and non-shaded regions.

Instance Level. The various current approaches for learning semantics in colorization, whether at the image or pixel levels, have limited ability to capture the variations in object appearance adequately. Although the previously reviewed models have demonstrated impressive results for a wide range of images, they experienced different problems when faced with images containing multiple objects on cluttered backgrounds. To solve such a problem, Su et al. [133] proposed the first instance-aware colorization architecture, incorporating an off-the-shelf object detector for capturing segmented object visuals as well as utilizing an instance-based colorization network to extract features at the object level. By leveraging this approach, the model could effectively handle complex scenes with multiple instances. Previous colorization methods have primarily focused on image-level or entity-level features, thereby not adequately capturing how the object instances in an image interact within the overall context. To address this limitation, Pucci et al. [134] introduced UCapsNet with capsules [127], incorporating features at the image level, produced by convolutions, and features at the entity level. Xia et al. [135] introduced the concept of disentangled colorization and proposed the identification of multiple color anchors that can effectively represent the diverse color distribution within an image. By specifying these anchor colors, their algorithm can predict the image colors, leveraging the global color affinity to ensure consistency in the overall structure. Chang et al. [87] built upon the concept of instance-awareness in colorization and introduced text guidance to establish a correspondence between instance regions and color descriptions. This allowed their model to assign colors to instances flexibly, even when such correspondence was not encountered during training.

In summary, colorization efforts benefit significantly from high-level image understanding, which has evolved towards finer granularity. Early methods treat the image as a whole, using image classification for global content understanding. This is referred to as the global level of understanding. Subsequently, methods evolved to the pixel level, where each pixel is classified through semantic segmentation. More recently, instance-level understanding has emerged, allowing different color assignments for different instances.

2.2.3 Applications for Vision Tasks

Unconditional image colorization technology can also aid other vision tasks. By illustrating two use cases, computational imaging as well as image restoration and enhancement, we demonstrate to readers that the research perspective on colorization technology extends beyond developing novel coloring algorithms. It also involves leveraging colorization technology to support other areas of research.

Computational Imaging. To improve the imaging quality of monochrome-color dual-lens systems, Dong et al. [136] tackled the reference-based colorization problem by expanding the number of pixels in the reference image and determining the color of each pixel in the target image as the ranked average of the colors. They used ResNet to capture high-level features from the input and reference images, refining the intermediate features and establishing feature relationships using convolution and Sigmoid layers. As an extension of this work, Dong et al. [137], leveraging a colorization CNN, introduced two parallel modules for colorization and colorization quality estimation. This work further advanced techniques for improving the colorization process and had great practical value for improving the imaging quality of mobile devices in various applications.

Image Restoration and Enhancement. In the computer vision field, some researchers proposed a general and multi-purpose vision model for several low-level image processing tasks and often evaluated the model performance in many applications, including colorization. The seminal work by Ulyanov et al. [138] demonstrated that a randomly initialized CNN implicitly captured texture-level image priors when trained iteratively on a single image, indicating that the CNN can be fine-tuned for image restoration tasks, such as reconstructing a corrupted image. Building on this, Pan et al. [139] expanded upon existing methods by utilizing image priors learned from a GAN trained on extensive natural images. This approach enabled flexible restoration and manipulation, unlike the fixed generator assumptions in previous GAN inversion techniques. Their method exhibited strong generalization capabilities across various image restoration and manipulation tasks despite not being tailored to each specific task. It restored missing information while preserving semantic details when reconstructing corrupted images. This demonstrated the potential for leveraging deep learning models to capture

and apply rich image priors to diverse image restoration tasks. Zhang et al. [140], using a unified model, introduced an efficient SCSNet paradigm that used low-resolution gray-scale images as input and produced high-resolution colorful images as outputs. Their model performed colorization and super-resolution in two consecutive stages, the first stage incorporating a Pyramid Valve Cross Attention (PVCAttn) module to combine information of the source and reference images effectively. A Continuous Pixel Mapping (CPM) module was developed in the second stage to efficiently generate target images at any magnification, using discrete pixel correlations in a continuous space. El Helou and Süsstrunk [141] viewed colorization as a generalization of classic restoration algorithms, such as image colorization, inpainting, and denoising. They decoupled the image restoration task into prior and data fidelity using an inversion GAN model, where a prior represented added information, such as the missing two-channel information for colorization. Wang et al. [142] introduced a zero-shot framework designed for arbitrary linear image restoration problems, offering a versatile solution for diverse image restoration endeavors, including image super-resolution, colorization, inpainting, compressed sensing, and deblurring. In a related study, Chan et al. [143] investigated the latent bank of GAN to uncover existing natural image priors. They proposed a novel approach that utilized pre-trained GAN models for various image restoration tasks, including super-resolution, colorization, and hybrid restoration.

2.3 Video Colorization

Although video colorization technology can be classified similarly to image techniques into conditional and unconditional methods, as an extension of image technology, it has unique challenges, including addressing video-specific user guidance and handling temporal information. We discuss video colorization techniques separately in the taxonomy to highlight its uniqueness.

Yatziv and Sapiro [144] proposed an approach that utilized an intrinsic, gradient-weighted distance measure to spread user-provided scribble colors across entire images or image sequences. Similarly, Jacob and Gupta [145] also developed a method that relied heavily on user inputs regarding segmentation and colorization. Their method improved the process by using a keyframe to transfer color to other frames, employing a motion estimation algorithm. Sheng et al. [146] introduced a different approach that involved establishing pixel similarity in the video's gray-scale channel, enabling parallel color optimization among pixels across video frames.

In the present deep learning era, we can broadly categorize existing video colorization methods into four types. The first type involves directly applying colorization over a single image and then post-processing the colorized image to achieve temporal consistency, as described previously in [147]. This method ensures that the colorization remains consistent over time, thereby providing a more realistic output. Another study [148] presented a specific case that demonstrated that training a model with Deep Video Prior can directly produce temporal consistency.

The second type of method performs colorization on individual frames and encodes temporal consistency by using motion estimation [149], [150] or inter-frame similarity using diffusion priors [151], [152]. This approach ensures that the colorization remains consistent with the inter-frame correspondence throughout the video sequence. Nevertheless, errors can accumulate due to inaccurate correspondence estimation.

The third type of method follows the tradition of example-based colorization, a colored frame being used as a reference, with the colorization of subsequent frames being achieved by evaluating content similarity [153], [154], [155], [156], [157], [158], [159], [160]. This method ensures a consistent color scheme throughout the video, based on the reference frame. Additionally, other methods within this type employed information propagation techniques to color the video frames [161], [162], [163], [164], [165], [166]. This approach allows for transferring color information from one frame to another, ensuring coherent colorization across the video.

The fourth and last type of method considers spatiotemporal features directly [167], [168], which eliminates the need for motion estimation or sequential information transfer from frame to frame, making it a more efficient method of video colorization.

3 Datasets

There are two situations in which researchers use datasets for colorization purposes: one for testing the performance of traditional methods, where researchers search the website for images, and the other for training networks, using large-scale image datasets, when designing deep learning-based models. This survey reviews the datasets used for **training** learning-based colorization models, as captured in Table 1. Datasets have been sorted chronologically according to the publication date and classified according to the types of input to be colorized. ImageNet [170] is the most widely used large-scale dataset for natural gray-scale image colorization, and researchers use this dataset to train models so as to acquire the color prior. Most of these gray-scale image datasets are mainly used for other vision tasks, although they are also suitable for colorization tasks due to their detailed annotations and scene categories. Researchers can obtain many data pairs by converting the original color images into gray-scale images. For video colorization, researchers mainly use DAVIS [180] and Videvo [182] datasets to train their models, especially for learning spatiotemporal consistency. Unlike natural images which can be captured easily on a large scale, collecting line drawing data presents a challenge, essentially being two ways of doing so: the first is to invite professional painters to draw manually, which is relatively expensive and time-consuming, and the second way is to extract the lines from colored animation data using SketchKeras [186] or Anime2Sketch [187], although the latter generally has line types which are very different to actual hand-painted line drawings. In addition to the above, the Danbooru dataset [184] is also a valuable resource for anyone interested in anime-style artwork. The most notable feature is its extensive tagging system. Images are annotated with a wide range of tags describing various aspects of the content, such as characters, themes, styles, and specific elements. This rich tagging system makes the

TABLE 1
An Overview of Major Datasets Used for Training Networks in Colorization Tasks

Category	Dataset	Year	Type	References
Gray-scale Image	CIFAR [169]	2009	Image classification	[108], [109]
	ImageNet [170]	2009	Visual recognition	[11], [39], [43], [44], [48], [49], [62], [64], [65], [107] [100], [103], [104], [105], [108], [109], [157], [160], [163]
	PASCAL VOC [171]	2010	Object recognition	[46], [130], [131]
	COCO [172]	2014	Object understanding	[43], [47], [81]
	Places 205 [173]	2014	Scene recognition	[46], [125], [129]
	LSUN [174]	2015	Scene understanding	[104], [105], [116]
	ADE20K [175]	2017	Semantic segmentation	[40]
Color Platte	MIT-Adobe 5K [176]	2011	Image enhancement	[177]
	Color Theme [78]	2012	Image colorization	[78]
Text Guidance	CoSaL [82]	2018	Image colorization	[82]
	Palette-and-Text [79]	2018	Image colorization	[79]
	SketchyScene [178]	2018	Sketch colorization	[179]
	Extended COCO-stuff [86]	2022	Image colorization	[84], [86], [87], [88]
	Multi-instance [87]	2023	Image colorization	[87], [88]
Video	DAVIS [180]	2016	Video object segmentation	[150], [154], [157], [160], [162], [165]
	Kinetics [181]	2017	Human action recognition	[156]
	Videvo [182]	2018	Free stock video footage	[150], [154], [160], [163], [165]
Line Drawing	Manga109 [183]	2017	Media processing of manga	[18]
	Danbooru2018 [184]	2018	Anime character recognition	[21], [53], [61], [70], [73], [74], [75], [76]
	Anime Sketch Colorization Pair [185]	2019	Sketch colorization	[19], [24], [57]
	Tag2Pix [85]	2019	Sketch colorization	[55], [77], [85]

dataset particularly valuable for training machine learning models in tasks like image classification, object detection, style transfer, and colorization.

4 Colorization Assessment Method

4.1 Limitations of Existing Evaluation Metrics

Assessing the output quality of various colorization methods is inherently linked to human visual aesthetic perception. A hybrid evaluation methodology, incorporating subjective and objective assessments, is standard in the published literature. Subjective evaluation involves qualitative analysis provided by the model designers and user studies based on participants' visual perceptions. Commonly used objective evaluation metrics for assessing colorization performance include Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), Learned Perceptual Image Patch Similarity (LPIPS), and Frechet Inception Distance (FID). Huang et al. [14] provided a detailed analysis of the existing evaluation metrics used in colorization. Nevertheless, these first three evaluation metrics were initially designed for image quality assessment [190] rather than for colorization quality assessment. They evaluate the colorization quality by calculating the difference between the colored image and the original color image of the input gray-scale image, i.e., the generated image should be as close to the original colored image as possible. Given that the colored rendition of a gray-scale image lacks uniqueness, achieving an exact replication of the original color image is deemed unnecessary. There should not be a unique output for any given input gray-scale image. Most learning-based methods train the colorization networks with paired training data. As a result, the mainstream evaluation method compares the colorization results with ground-truth color images in the test dataset. In addition, FID is a metric used to evaluate the performance of generative models. Although some generative model-based colorization methods may exhibit superior generation performance, this does not necessarily mean that their colorization performance aligns with the generation abilities.

4.2 Colorization Aesthetic Assessment

In this paper, in addition to reviewing the current state-of-the-science, we explore the use of a neural network model to simulate human visual perception to evaluate the performance of **automatic colorization models.** Inspired by the work of image aesthetic assessment [191], we propose a new concept of *colorization aesthetic assessment.* Human aesthetic evaluation of colors involves psychology and visual cognition elements, which hold significant practical value in computational imaging. For example, when developing imaging algorithms for mobile phone photography, different cell phone manufacturers apply distinctly different color adjustments, resulting in different color effects in photos of the same scene. The key problem of aesthetic evaluation is quantifying subjective judgment. Hence, we use two metrics based on large-scale cross-modal data training to conduct our coloring aesthetic evaluation test. In this survey, we mainly focus on testing the unconditional natural gray-scale colorization performance since the automatic coloring method can reflect the model's ability for semantic understanding and color representation. For demonstration purposes, the two metrics described below are used to evaluate the following seven representative colorization models: Colorful Colorization [100], User-Guided Colorization [62], InstColor [133], BigColor [121], UniColor [93], Disco [135], and DDColor [113]. All these models use automatic colorization methods or are set in automatic mode, which means no user-guided color information is included.

The first metric, CLIP-IQA [188], represents an image quality assessment metric, leveraging the visual language prior embedded in the CLIP model [94]. This metric enables the evaluation of both the perceptual quality and abstract perception of images without the need for explicit task-specific training. Rather than computing the cosine similarity between two feature vectors of image and text, the cosine similarity between the image feature vector x and every antonym prompt pair (t_1, t_2) is instead first computed

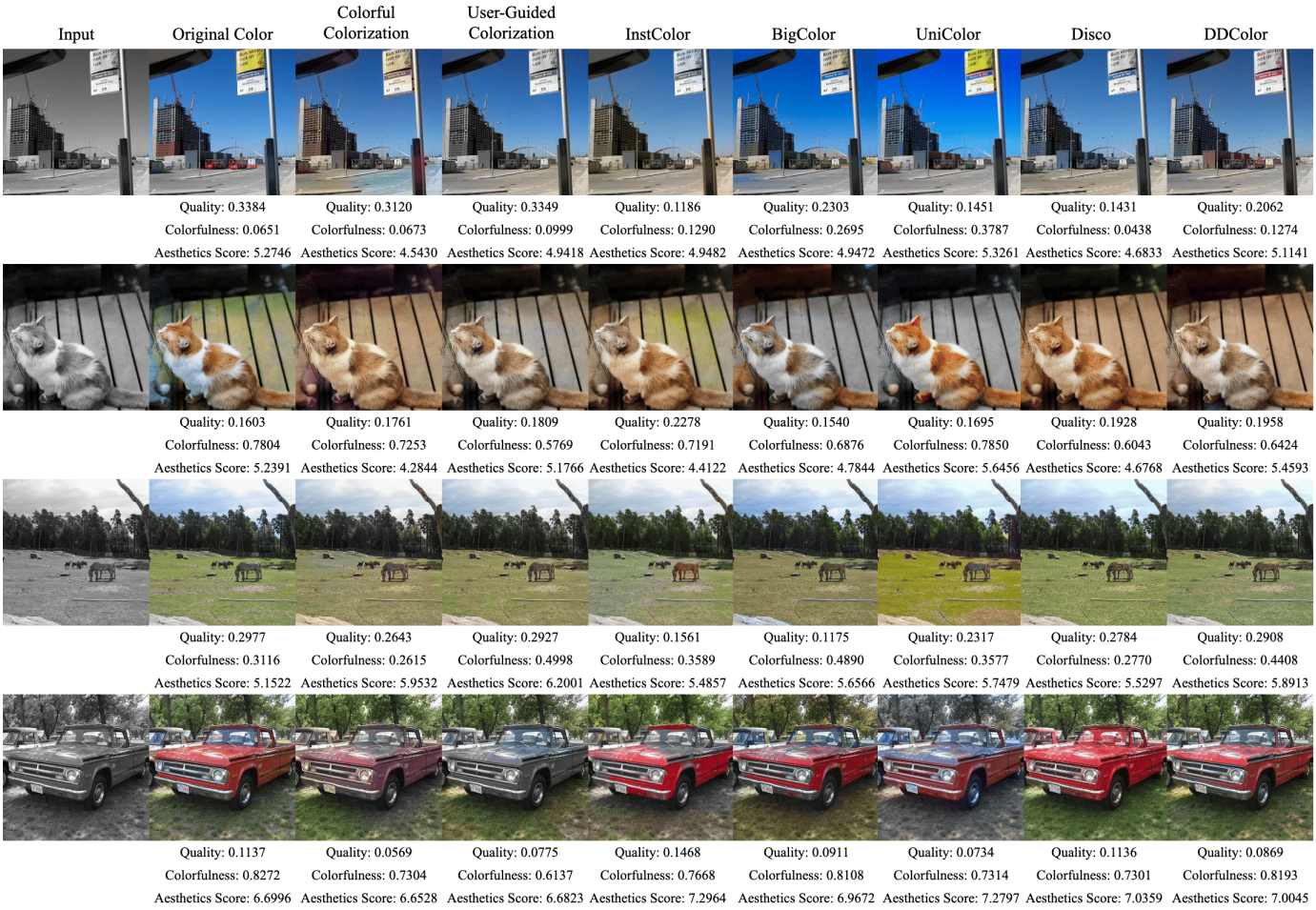

Fig. 9. Colorization aesthetic assessment of seven automatic colorization models, including Colorful Colorization [100], User-Guided Colorization [62], InstColor [133], BigColor [121], UniColor [93], Disco [135], and DDColor [113]. Under each example of the original color image and the colored image, we show three scores computed by CLIP-IQA [188] and LAION-Aesthetics Predictor V2 [189]. The first two scores provide a measure of the image quality on the basis of the image's overall texture and color dimensions, while the last score represents the image's aesthetic quality. It can be seen that the colorization results of some samples exceed the original color images in terms of their aesthetic scores.

to reduce the ambiguity of text prompts as follows:

$$s_i = \frac{x \bigodot t_i}{\|x\| \cdot \|t_i\|}, i \in \{1, 2\}, \tag{3}$$

Softmax is then used to compute the final score $\overline{s} \in [0, 1]$ as follows:

$$\overline{s} = \frac{e^{s_1}}{e^{s_1} + e^{s_2}}. \tag{4}$$

The CLIP-IQA metric can assess overall and fine-grained image quality and contains many built-in evaluation prompts. We mainly use 'quality' and 'colorfulness' to evaluate colorized images, including quality and color perceptions. Here, 'quality' refers to the overall picture quality of the image, and 'colorfulness' measures how rich and saturated the colors appear. The higher the score, the higher the image quality and contribution to the image's overall visual impact and aesthetic appeal.

The second metric, LAION-Aesthetics Predictor V2 [189], is an image aesthetic assessment metric that takes CLIP image embeddings produced using the clip-vit-large-patch14 model as input and then outputs the aesthetic score, ranging from 1 to 10, by concatenating a simple linear model. It is trained on SAC, LAION-Logos, and AVA datasets. The higher the score, the higher the aesthetic quality.

4.3 Experiments

We performed non-reference image quality and aesthetic assessments of seven representative colorization methods, using COCO test dataset [172] with central cropped testing images of resolution 256×256. As illustrated in Fig. 9, these seven colorization methods accurately represent the global semantic analysis of gray-scale images. The color prior, learned by means of the training on a large number of natural image datasets, can identify the sky as blue and the lawn as green. Nevertheless, the automatic colorization method is unsuitable for controlling all the details, such as in the case of the bus in the top row of Fig. 9. An interactive approach would work better for such a fine-grained coloring task. In addition, the examples in the second and fourth rows of Fig. 9 also reveal pattern collapse problems when training with paired data. The seven methods tend to consider the patterned fur coat of the cat to be orange and the car's color to be red.

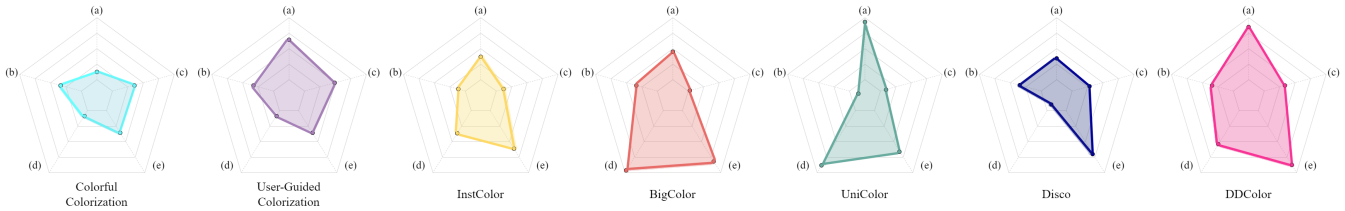

Fig. 10. Visualization of the performance of seven automatic colorization methods, including Colorful Colorization [100], User-Guided Colorization [62], InstColor [133], BigColor [121], UniColor [93], Disco [135], and DDColor [113]. (a) Aesthetic Score, (b) Inference Time, (c) Quality, (d) Colorfulness, and (e) FID. All values are the normalized average scores.

TABLE 2
Quantitative Comparison of Seven Automatic Colorization Methods Using Existing Evaluation Metrics

Method	COCO test dataset (5k)				ImageNet test dataset (10k)			
	PSNR↑	SSIM↑	LPIPS↓	FID↓	PSNR↑	SSIM↑	LPIPS↓	FID↓
Colorful Colorization [100]	22.96	**0.998**	0.213	17.90	22.84	**0.998**	0.219	9.72
User-Guided Colorization [62]	**24.75**	**0.998**	**0.171**	16.82	**24.67**	**0.998**	**0.180**	8.55
InstColor [133]	22.35	0.838	0.238	12.72	22.03	0.909	0.217	7.35
BigColor [121]	21.40	0.887	0.214	9.05	21.57	0.886	0.212	**3.67**
UniColor [93]	23.09	0.912	0.202	11.16	21.73	0.909	0.236	6.93
Disco [135]	20.46	0.851	0.236	10.59	20.72	0.862	0.229	5.57
DDColor [113]	23.41	0.997	0.184	**7.88**	23.18	0.997	0.192	3.93

TABLE 3
Inference Time Comparison of Eight Automatic Colorization Methods

Method	Time cost (seconds)
Colorful Colorization [100]	**0.0749**
User-Guided Colorization [62]	0.0853
InstColor [133]	0.8144
BigColor [121]	0.0875
UniColor [93]	1.7162
Disco [135]	0.0945
DDColor [113]	0.0779
iColoriT [65]	0.1121

Regarding aesthetic evaluation, UniColor [93] produces coloring results of high saturation and achieves the highest scores in terms of color and aesthetics. BigColor [121] can also produce a high-saturation coloring effect. In particular, the cat example displays a prominent cat body coloring effect, which is even more beautiful than the original image. InstColor [133], Disco [135], and DDColor [113] all produce similar color effects, and which is relatively soft. In terms of score, DDColor performs slightly better than the other two. Colorful Colorization [100] and User-Guided Colorization [62] are two early learning-based methods. Their results exhibit problems of color bleeding (the third column of the top row) or incomplete coloring (the fourth column of the bottom row) on the individual examples. According to the experimental data, the quality score is not always positively correlated to the aesthetic scores, with some samples having high aesthetic scores but low-quality scores. This is due to coloring being a process of image recovery and information enhancement of gray-scale images, and different methods have different processing mechanisms for gray-scale images. Hence, the information gain and image loss are different. As is illustrated in Fig. 10, we used a radar-type map to visualize indicators in five dimensions, including (a) Aesthetic Score, (b) Inference Time, (c) Quality, (d) Colorfulness, and (e) FID. All the values have been normalized, and the larger the area, the better the overall performance of the particular method.

We also evaluated the automatic colorization methods using traditional assessment metrics on the COCO test and ImageNet test dataset, the results being captured in Table 2. Bold indicates the best, and underline indicates the next best. To maintain the object structure without geometry distortion, all images have been centrally cropped to a resolution of 256×256 instead of directly resizing. The reconstruction-based evaluation metrics and generation capability index does not intuitively reflect the performance of the colorization models for color information representation. In the case of both datasets, the FID scores illustrate that the early methods [62], [100], which do not use generative models, perform worse than the other methods for the diversity of data samples. In Table 3, we compare the inference time of the eight automatic models, all tested on a single RTX 4090 GPU, with the values in the table representing the average time taken to colorize each image. In addition to the above-discussed seven methods, we demonstrate another method, iColoriT [65], which can also perform unconditional colorization, although it is mainly designed for interactively color-hint solutions. It is worth noting that iColoriT is mainly designed for user hint interactive colorization method; when it is used for unconditional colorization, this so-called automatic mode is two-staged, i.e., firstly use their provided source code to generate randomly sampling hint locations, and then according to the generated color hints to perform colorization. In this experiment, the generated color-hints are given by the ImageNet ctest 10k dataset, so we don't need to include the time cost of color-hint generation. After a comprehensive evaluation, DDColor [113] was found to be the current best automatic gray-scale image colorization method. Using the

theoretical aspects of vision to evaluate the images produced by graphics is very much in line with the research concept of 'vision for graphics,' such as assessing the generated images of Artificial Intelligence-Generated Content (AIGC) models. We hope our initial attempts to perform a colorization aesthetic assessment will inspire the research community to pursue such an aesthetic-based approach and understanding in their future research. This would significantly simplify and reduce the labor-intensive and burdensome nature of colorization in the future.

5 Discussion and Future Work

Colorization technology has gradually developed over two decades, from the initial user-guided gray-scale image rendering technology to the more advanced learning-based generative multi-modal colorization. The coloring targets also varied during the period, covering static images and videos in gray-scale or line art. In this section, we suggest potentially valuable research directions in colorization that need to be explored.

Integrate colorization with AIGC technology. The explosion of AIGC technology has revolutionized colorization, reforming the reference-based image colorization task as an exemplar-guided image editing task. In particular, the field of animation content generation shows significant research potential. ControlNet-based work [92] demonstrates the excellent performance of generative models in sketch-guided anime content generation. With generative algorithms [5], [7], [8], tasks such as sketch frame interpolation, reference-based colorization, and animation generation can be accomplished using a single model, significantly enhancing animators' work efficiency. As a valuable topic for future research, we propose designing algorithms that can support the coloring of line drafts with large movements and address the issue of poor coloring results during screen transitions, as existing methods do not perform well in these scenarios.

The trend of multi-modal interaction methods. At the early development stage of colorization technology, methods based on user guidance were proposed, with various user input forms. Nevertheless, using a single form is always inadequate. Hence, multi-modal fusion is considered worthy of further in-depth research. These pioneering multi-modal methods [93], [95], [96] verify the effectiveness of the multi-modal coloring approach. Human guidance and interaction are still essential in the coloring process, and the neural network model only helps us fill in some color priors. Within this context, various details must be manually fine-tuned to generate final plausible results.

Generalization ability of reference-based methods. The reference-based methods, especially for line drawing colorization, have been extensively studied [19], [21], [22], [54], [56], in which cross-domain feature matching is the core problem that always needs to be solved. Since the lack of paired training data, Lee et al. [54] developed a self-supervised augmented training strategy as a compromise, formulating the problem as a line drawing guided image restoration task to learn the cross-domain feature matching. However, this method's generalization performance is relatively poor. Different from real-world images, the feature correspondence of cartoons needs to be designed separately and trained through customized datasets.

Image and Video Editing. Colorization techniques can be used for image and video editing, such as recolorization or color transfer [192], [193]. In addition to the editing of color information in photographs and videos, researchers are increasingly interested in more complex tasks, such as non-photorealistic artistic rendering of images [194] or videos [195], [196]. In particular, with the help of text control, various artistic edits can be carried out on the screen content, which has strong feasibility and commercial value. As an extension to traditional coloring techniques, an essential research direction should be executing these 'magic applications' with low power consumption so that users can realize these operations more conveniently on a mobile phone.

Dedicated new assessment methods for colorization. To evaluate the performance of colorization methods, researchers have utilized different assessment methods, including subjective and objective evaluation metrics. Huang et al. [14] have detailed the traditional evaluation metrics used in research. In this paper, we also initially attempt to introduce aesthetic assessment into colorization, and it is still an open and challenging problem to design an evaluation system tailored for colorization. As the assessment of coloring results is highly subjective, and aesthetics is one of the key aspects, developing a more accurate and comprehensive coloring evaluation system would be a valuable research direction.

6 Concluding Remarks

This survey provides an overview of the research and state-of-the-science in the field of colorization technology, which originated from application research in computer graphics. Based on this view, we provide a taxonomy of colorization technology and describe the subcategories and contents. Our systematic investigation and analysis conclude that colorization originates in graphics, excels in vision, and forms a fusion with the rapid development of generative models (i.e., *vision for graphics*). In the process, we also extend current colorization evaluation metrics and propose a concept of colorization aesthetic assessment for evaluating seven automatic colorization methods. We also explore the challenges and potential future research directions in colorization. Finally, we hope this survey serves as a valuable resource and inspiration for future researchers in colorization.

Acknowledgements

The work was partially supported by grants from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. 15602323) and The Hong Kong Polytechnic University (Project No. P0049355/CD95). The National Science and Technology Council also partly funded this research under Grant 113-2221-E-006-161-MY3 and 111-2221-E-006-112-MY3, Taiwan.

REFERENCES

[1] R. Xu, Z. Tu, Y. Du, X. Dong, J. Li, Z. Meng, J. Ma, A. Bovik, and H. Yu, "Pik-Fix: restoring and colorizing old photos," in *IEEE Winter Conference on Applications of Computer Vision*, 2023, pp. 1724–1734.

[2] S. Iizuka and E. Simo-Serra, "Deepremaster: temporal source-reference attention networks for comprehensive video enhancement," *ACM Transactions on Graphics*, vol. 38, no. 6, pp. 176:1–176:13, 2019.

[3] D. Varga, C. A. Szabó, and T. Szirányi, "Automatic cartoon colorization based on convolutional neural network," in *International Workshop on Content-Based Multimedia Indexing*, 2017, pp. 28:1–28:6.

[4] X. Li, B. Zhang, J. Liao, and P. V. Sander, "Deep sketch-guided cartoon video inbetweening," *IEEE Transactions on Visualization and Computer Graphics*, vol. 28, no. 8, pp. 2938–2952, 2022.

[5] J. Xing, H. Liu, M. Xia, Y. Zhang, X. Wang, Y. Shan, and T.-T. Wong, "ToonCrafter: generative cartoon interpolation," *arXiv preprint arXiv:2405.17933*, pp. 1–12, 2024.

[6] Y. Dai, S. Zhou, Q. Li, C. Li, and C. C. Loy, "Learning inclusion matching for animation paint bucket colorization," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2024, pp. 25 544–25 553.

[7] Z. Huang, M. Zhang, and J. Liao, "LVCD: reference-based lineart video colorization with diffusion models," *ACM Transactions on Graphics*, vol. 43, no. 6, pp. 177:1–177:11, 2024.

[8] Y. Meng, H. Ouyang, H. Wang, Q. Wang, W. Wang, K. L. Cheng, Z. Liu, Y. Shen, and H. Qu, "AniDoc: animation creation made easier," *arXiv preprint arXiv:2412.14173*, pp. 1–11, 2024.

[9] J. R. Truckenbrod, "Effective use of color in computer graphics," in *ACM SIGGRAPH*, vol. 15, no. 3, 1981, p. 83–90.

[10] G. Charpiat, M. Hofmann, and B. Schölkopf, "Automatic image colorization via multimodal predictions," in *European Conference on Computer Vision*, 2008, pp. 126–139.

[11] W. Yin, P. Lu, Z. Zhao, and X. Peng, "Yes, "attention is all you need", for exemplar based colorization," in *ACM International Conference on Multimedia*, 2021, pp. 2243–2251.

[12] S.-Y. Chen, J.-Q. Zhang, Y.-Y. Zhao, P. L. Rosin, Y.-K. Lai, and L. Gao, "A review of image and video colorization: From analogies to deep learning," *Visual Informatics*, vol. 6, no. 3, pp. 51–68, 2022.

[13] S. Anwar, M. Tahir, C. Li, A. Mian, F. S. Khan, and A. W. Muzaffar, "Image colorization: A survey and dataset," *Information Fusion*, pp. 1–31, 2024.

[14] S. Huang, X. Jin, Q. Jiang, and L. Liu, "Deep learning for image colorization: Current and future prospects," *Engineering Applications of Artificial Intelligence*, vol. 114, no. C, pp. 1–27, 2022.

[15] T. Welsh, M. Ashikhmin, and K. Mueller, "Transferring color to greyscale images," *ACM Transactions on Graphics*, vol. 21, no. 3, pp. 277–280, 2002.

[16] A. Levin, D. Lischinski, and Y. Weiss, "Colorization using optimization," *ACM Transactions on Graphics*, vol. 23, no. 3, pp. 689–694, 2004.

[17] Y. Qu, T.-T. Wong, and P.-A. Heng, "Manga colorization," *ACM Transactions on Graphics*, vol. 25, no. 3, pp. 1214–1220, 2006.

[18] C. Furusawa, K. Hiroshiba, K. Ogaki, and Y. Odagiri, "Comicolorization: Semi-automatic manga colorization," in *ACM SIGGRAPH Asia Technical Briefs*, 2017, pp. 12:1–12:4.

[19] Y. Cao, H. Tian, and P. Y. Mok, "Attention-aware anime line drawing colorization," in *IEEE International Conference on Multimedia and Expo*, 2023, pp. 1637–1642.

[20] D. Yan, R. Ito, R. Moriai, and S. Saito, "Two-Step Training: adjustable sketch colourization via reference image and text tag," *Computer Graphics Forum*, vol. 42, no. 6, pp. 1–14, 2023.

[21] Y. Cao, X. Meng, P. Y. Mok, T.-Y. Lee, X. Liu, and P. Li, "AnimeDiffusion: Anime diffusion colorization," *IEEE Transactions on Visualization and Computer Graphics*, vol. 30, no. 10, pp. 6956–6969, 2024.

[22] J. Lin, W. Zhao, and Y. Wang, "Visual correspondence learning and spatially attentive synthesis via transformer for exemplar-based anime line art colorization," *IEEE Transactions on Multimedia*, vol. 26, pp. 6880–6890, 2024.

[23] Y. Liu, Z. Qin, T. Wan, and Z. Luo, "Auto-painter: Cartoon image generation from sketch by using conditional wasserstein generative adversarial networks," *Neurocomputing*, vol. 311, pp. 78–87, 2018.

[24] X. Liu, W. Wu, C. Li, Y. Li, and H. Wu, "Reference-guided structure-aware deep sketch colorization for cartoons," *Computational Visual Media*, vol. 8, no. 1, pp. 135–148, 2022.

[25] S.-Y. Chen, J.-Q. Zhang, L. Gao, Y. He, S. Xia, M. Shi, and F.-L. Zhang, "Active colorization for cartoon line drawings," *IEEE Transactions on Visualization and Computer Graphics*, vol. 28, no. 2, pp. 1198–1208, 2022.

[26] T.-H. Sun, C.-H. Lai, S.-K. Wong, and Y.-S. Wang, "Adversarial colorization of icons based on contour and color conditions," in *ACM International Conference on Multimedia*, 2019, p. 683–691.

[27] Y. Li, Y. Lien, and Y. Wang, "Style-structure disentangled features and normalizing flows for diverse icon colorization," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11 234–11 243.

[28] A. Hertzmann, C. E. Jacobs, N. Oliver, B. Curless, and D. H. Salesin, "Image analogies," in *ACM SIGGRAPH*, 2001, pp. 327–340.

[29] R. Irony, D. Cohen-Or, and D. Lischinski, "Colorization by example," in *Eurographics Symposium on Rendering*, 2005, pp. 201–210.

[30] X. Liu, L. Wan, Y. Qu, T.-T. Wong, S. Lin, C.-S. Leung, and P.-A. Heng, "Intrinsic colorization," *ACM Transactions on Graphics*, vol. 27, no. 5, pp. 152:1–152:9, 2008.

[31] Y. Morimoto, Y. Taguchi, and T. Naemura, "Automatic colorization of grayscale images using multiple images on the web," in *ACM SIGGRAPH: Talks*, 2009, p. 59:1.

[32] A. Y.-S. Chia, S. Zhuo, R. K. Gupta, Y.-W. Tai, S.-Y. Cho, P. Tan, and S. Lin, "Semantic colorization with internet images," *ACM Transactions on Graphics*, vol. 30, no. 6, pp. 1–8, 2011.

[33] R. K. Gupta, A. Y.-S. Chia, D. Rajan, E. S. Ng, and Z. Huang, "Image colorization using similar images," in *ACM International Conference on Multimedia*, 2012, pp. 369–378.

[34] A. Bugeau, V.-T. Ta, and N. Papadakis, "Variational exemplar-based image colorization," *IEEE Transactions on Image Processing*, vol. 23, no. 1, pp. 298–307, 2014.

[35] B. Li, F. Zhao, Z. Su, X. Liang, Y.-K. Lai, and P. L. Rosin, "Example-based image colorization using locality consistent sparse representation," *IEEE Transactions on Image Processing*, vol. 26, no. 11, pp. 5188–5202, 2017.

[36] B. Li, Y.-K. Lai, M. John, and P. L. Rosin, "Automatic example-based image colorization using location-aware cross-scale matching," *IEEE Transactions on Image Processing*, vol. 28, no. 9, pp. 4606–4619, 2019.

[37] F. Fang, T. Wang, T. Zeng, and G. Zhang, "A superpixel-based variational model for image colorization," *IEEE Transactions on Visualization and Computer Graphics*, vol. 26, no. 10, pp. 2931–2943, 2020.

[38] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, vol. 25, 2012, pp. 1097–1105.

[39] M. He, D. Chen, J. Liao, P. V. Sander, and L. Yuan, "Deep exemplar-based colorization," *ACM Transactions on Graphics*, vol. 37, no. 4, pp. 47:1–47:16, 2018.

[40] C. Xiao, C. Han, Z. Zhang, J. Qin, T.-T. Wong, G. Han, and S. He, "Example-based colourization via dense encoding pyramids," *Computer Graphics Forum*, vol. 39, no. 1, pp. 20–33, 2020.

[41] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *IEEE International Conference on Computer Vision*, 2017, pp. 1510–1519.

[42] Y. Li, M.-Y. Liu, X. Li, M.-H. Yang, and J. Kautz, "A closed-form solution to photorealistic image stylization," in *European Conference on Computer Vision*, 2018, pp. 468–483.

[43] Z. Xu, T. Wang, F. Fang, Y. Sheng, and G. Zhang, "Stylization-based architecture for fast deep exemplar colorization," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9360–9369.

[44] P. Lu, J. Yu, X. Peng, Z. Zhao, and X. Wang, "Gray2ColorNet: Transfer more colors from reference image," in *ACM International Conference on Multimedia*, 2020, pp. 3210–3218.

[45] C. L. P. Chen and Z. Liu, "Broad learning system: An effective and efficient incremental learning system without the need for deep architecture," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 1, pp. 10–24, 2018.

[46] H. Li, B. Sheng, P. Li, R. Ali, and C. L. P. Chen, "Globally and locally semantic colorization via exemplar-based Broad-GAN," *IEEE Transactions on Image Processing*, vol. 30, pp. 8526–8539, 2021.

[47] H. Carrillo, M. Clément, and A. Bugeau, "Super-attention for exemplar-based image colorization," in *Asian Conference on Computer Vision*, 2022, pp. 646–662.

[48] Y. Bai, C. Dong, Z. Chai, A. Wang, Z. Xu, and C. Yuan, "Semantic-sparse colorization network for deep exemplar-based colorization," in *European Conference on Computer Vision*, 2022, pp. 505–521.

[49] H. Wang, D. Zhai, X. Liu, J. Jiang, and W. Gao, "Unsupervised deep exemplar colorization via pyramid dual non-local attention," *IEEE Transactions on Image Processing*, vol. 32, pp. 4114–4127, 2023.

[50] L. Zhang, Y. Ji, X. Lin, and C. Liu, "Style transfer for anime sketches with enhanced residual u-net and auxiliary classifier GAN," in *IAPR Asian Conference on Pattern Recognition*, 2017, pp. 506–511.

[51] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2414–2423.

[52] A. Odena, C. Olah, and J. Shlens, "Conditional image synthesis with auxiliary classifier GANs," in *International Conference on Machine Learning*, vol. 70, 2017, pp. 2642–2651.

[53] K. Akita, Y. Morimoto, and R. Tsuruno, "Colorization of line drawings with empty pupils," *Computer Graphics Forum*, vol. 39, no. 7, pp. 601–610, 2020.

[54] J. Lee, E. Kim, Y. Lee, D. Kim, J. Chang, and J. Choo, "Reference-based sketch image colorization using augmented-self reference and dense semantic correspondence," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5800–5809.

[55] R. Cao, H. Mo, and C. Gao, "Line art colorization based on explicit region segmentation," *Computer Graphics Forum*, vol. 40, no. 7, pp. 1–10, 2021.

[56] Z. Li, Z. Geng, Z. Kang, W. Chen, and Y. Yang, "Eliminating gradient conflict in reference-based line-art colorization," in *European Conference on Computer Vision*, 2022, pp. 579–596.

[57] S. Wu, X. Yan, W. Liu, S. Xu, and S. Zhang, "Self-driven dual-path learning for reference-based line art colorization under limited data," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–15, 2023.

[58] Y.-C. Huang, Y.-S. Tung, J.-C. Chen, S.-W. Wang, and J.-L. Wu, "An adaptive edge detection based colorization algorithm and its applications," in *ACM International Conference on Multimedia*, 2005, pp. 351–354.

[59] L. Yatziv and G. Sapiro, "Fast image and video colorization using chrominance blending," *IEEE Transactions on Image Processing*, vol. 15, no. 5, pp. 1120–1129, 2006.

[60] Q. Luan, F. Wen, D. Cohen-Or, L. Liang, Y.-Q. Xu, and H.-Y. Shum, "Natural image colorization," in *Eurographics Symposium on Rendering*, 2007, pp. 309–320.

[61] L. Zhang, C. Li, T.-T. Wong, Y. Ji, and C. Liu, "Two-stage sketch colorization," *ACM Transactions on Graphics*, vol. 37, no. 6, pp. 261:1–261:14, 2018.

[62] R. Zhang, J.-Y. Zhu, P. Isola, X. Geng, A. S. Lin, T. Yu, and A. A. Efros, "Real-time user-guided image colorization with learned deep priors," *ACM Transactions on Graphics*, vol. 36, no. 4, pp. 119:1–119:11, 2017.

[63] E. Kim, S. Lee, J. Park, S. Choi, C. Seo, and J. Choo, "Deep edge-aware interactive colorization against color-bleeding effects," in *IEEE International Conference on Computer Vision*, 2021, pp. 14 647–14 656.

[64] Y. Xiao, J. Wu, J. Zhang, P. Zhou, Y. Zheng, C.-S. Leung, and L. Kavan, "Interactive deep colorization and its application for image compression," *IEEE Transactions on Visualization and Computer Graphics*, vol. 28, no. 3, pp. 1557–1572, 2022.

[65] J. Yun, S. Lee, M. Park, and J. Choo, "iColoriT: Towards propagating local hints to the right region in interactive colorization by leveraging vision transformer," in *IEEE Winter Conference on Applications of Computer Vision*, 2023, pp. 1787–1796.

[66] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021, pp. 1–21.

[67] D. Sýkora, J. Dingliana, and S. Collins, "Lazybrush: Flexible painting tool for hand-drawn cartoons," *Computer Graphics Forum*, vol. 28, no. 2, pp. 599–608, 2009.

[68] P. Sangkloy, J. Lu, C. Fang, F. Yu, and J. Hays, "Scribbler: Controlling deep image synthesis with sketch and color," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6836–6845.

[69] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, pp. 1–7, 2014.

[70] Y. Ci, X. Ma, Z. Wang, H. Li, and Z. Luo, "User-guided deep anime line art colorization with conditional adversarial networks," in *ACM International Conference on Multimedia*, 2018, pp. 1536–1544.

[71] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of wasserstein GANs," in *Advances in Neural Information Processing Systems*, vol. 30, 2017, pp. 5767–5777.

[72] J. Johnson, A. Alahi, and F.-F. Li, "Perceptual losses for real-time style transfer and super-resolution," in *European Conference on Computer Vision*, 2016, pp. 694–711.

[73] L. Zhang, C. Li, E. Simo-Serra, Y. Ji, T.-T. Wong, and C. Liu, "User-guided line art flat filling with split filling mechanism," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2021, pp. 9884–9893.

[74] M. Yuan and E. Simo-Serra, "Line art colorization with concatenated spatial attention," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2021, pp. 3941–3945.

[75] Z. Dou, N. Wang, B. Li, Z. Wang, H. Li, and B. Liu, "Dual color space guided sketch colorization," *IEEE Transactions on Image Processing*, vol. 30, pp. 7292–7304, 2021.

[76] H. Carrillo, M. Clément, A. Bugeau, and E. Simo-Serra, "Diffusart: Enhancing line art colorization with conditional diffusion models," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2023, pp. 3486–3490.

[77] Y. Cho, J. Lee, S. Yang, J. Kim, Y. Park, H. Lee, M. A. Khan, D. Kim, and J. Choo, "Guiding users to where to give color hints for efficient interactive sketch colorization via unsupervised region prioritization," in *IEEE Winter Conference on Applications of Computer Vision*, 2023, pp. 1818–1827.

[78] X.-H. Wang, J. Jia, H.-Y. Liao, and L.-H. Cai, "Affective image colorization," *Journal of Computer Science and Technology*, vol. 27, no. 6, pp. 1119–1128, 2012.

[79] H. Bahng, S. Yoo, W. Cho, D. K. Park, Z. Wu, X. Ma, and J. Choo, "Coloring with words: Guiding image colorization through text-based palette generation," in *European Conference on Computer Vision*, 2018, pp. 443–459.

[80] S. Wu, Y. Yang, S. Xu, W. Liu, X. Yan, and S. Zhang, "FlexIcon: flexible icon colorization via guided images and palettes," in *ACM International Conference on Multimedia*, 2023, p. 8662–8673.

[81] V. Manjunatha, M. Iyyer, J. Boyd-Graber, and L. Davis, "Learning to color from language," in *North American Chapter of the Association for Computational Linguistics*, 2018, pp. 764–769.

[82] J. Chen, Y. Shen, J. Gao, J. Liu, and X. Liu, "Language-based image editing with recurrent attentive models," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8721–8729.

[83] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, p. 1735–1780, 1997.

[84] Z. Chang, S. Weng, Y. Li, S. Li, and B. Shi, "L-CoDer: Language-based colorization with color-object decoupling transformer," in *European Conference on Computer Vision*, 2022, pp. 360–375.

[85] H. Kim, H. Y. Jhoo, E. Park, and S. Yoo, "Tag2Pix: Line art colorization using text tag with SECat and changing loss," in *IEEE International Conference on Computer Vision*, 2019, pp. 9055–9064.

[86] S. Weng, H. Wu, Z. Chang, J. Tang, S. Li, and B. Shi, "L-CoDe: Language-based colorization using color-object decoupled conditions," in *AAAI Conference on Artificial Intelligence*, vol. 36, no. 3, 2022, pp. 2677–2684.

[87] Z. Chang, S. Weng, P. Zhang, Y. Li, S. Li, and B. Shi, "L-CoIns: Language-based colorization with instance awareness," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2023, pp. 19 221–19 230.

[88] Z. Chang, S. Weng, P. Zhang, Y. Li, S. Li, and B. Shi, "L-CAD: Language-based colorization with any-level descriptions using diffusion priors," in *Advances in Neural Information Processing Systems*, vol. 36, 2023, pp. 77 174–77 186.

[89] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 674–10 685.

[90] H. Liu, J. Xing, M. Xie, C. Li, and T.-T. Wong, "Improved diffusion-based image colorization via piggybacked models," *arXiv preprint arXiv:2304.11105*, pp. 1–17, 2023.

[91] N. Zabari, A. Azulay, A. Gorkor, T. Halperin, and O. Fried, "Diffusing colors: Image colorization with text guided diffusion," in *ACM SIGGRAPH Asia*, 2023, pp. 61:1–61:11.

[92] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *IEEE International Conference on Computer Vision*, 2023, pp. 3836–3847.

[93] Z. Huang, N. Zhao, and J. Liao, "UniColor: A unified framework for multi-modal colorization with transformer," *ACM Transactions on Graphics*, vol. 41, no. 6, pp. 205:1–205:16, 2022.

[94] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*, vol. 139, 2021, pp. 8748–8763.

[95] Z. Liang, Z. Li, S. Zhou, C. Li, and C. C. Loy, "Control Color: multimodal diffusion-based interactive image colorization," *arXiv preprint arXiv:2402.10855*, pp. 1–10, 2024.

[96] D. Yan, L. Yuan, Y. Nishioka, I. Fujishiro, and S. Saito, "ColorizeDiffusion: adjustable sketch colorization with reference image and text," *arXiv preprint arXiv:2401.01456*, pp. 1–17, 2024.

[97] Z. Cheng, Q. Yang, and B. Sheng, "Deep colorization," in *IEEE International Conference on Computer Vision*, 2015, pp. 415–423.

[98] A. Deshpande, J. Rock, and D. Forsyth, "Learning large-scale automatic image colorization," in *IEEE International Conference on Computer Vision*, 2015, pp. 567–575.

[99] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[100] R. Zhang, P. Isola, and A. A. Efros, "Colorful image colorization," in *European Conference on Computer Vision*, 2016, pp. 649–666.

[101] G. Larsson, M. Maire, and G. Shakhnarovich, "Learning representations for automatic colorization," in *European Conference on Computer Vision*, 2016, pp. 577–593.

[102] G. Larsson, M. Maire, and G. Shakhnarovich, "Colorization as a proxy task for visual understanding," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 840–849.

[103] Y. Wu, X. Wang, Y. Li, H. Zhang, X. Zhao, and Y. Shan, "Towards vivid and diverse image colorization with generative color prior," in *IEEE International Conference on Computer Vision*, 2021, pp. 14 357–14 366.

[104] A. Deshpande, J. Lu, M.-C. Yeh, M. J. Chong, and D. Forsyth, "Learning diverse image colorization," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2877–2885.

[105] S. Messaoud, D. Forsyth, and A. G. Schwing, "Structural consistency and controllability for diverse colorization," in *European Conference on Computer Vision*, 2018, pp. 603–619.

[106] L. Ardizzone, C. Lüth, J. Kruse, C. Rother, and U. Köthe, "Guided image generation with conditional invertible neural networks," *arXiv preprint arXiv:1907.02392*, pp. 1–11, 2019.

[107] S. Guadarrama, R. Dahl, D. Bieber, M. Norouzi, J. Shlens, and K. Murphy, "PixColor: pixel recursive colorization," in *British Machine Vision Conference*, 2017, pp. 1–17.

[108] A. Royer, A. Kolesnikov, and C. H. Lampert, "Probabilistic image colorization," in *British Machine Vision Conference*, 2017, pp. 1–15.

[109] M. H. Baig and L. Torresani, "Multiple hypothesis colorization and its application to image compression," *Computer Vision and Image Understanding*, vol. 164, pp. 111–123, 2017.

[110] M. Kumar, D. Weissenborn, and N. Kalchbrenner, "Colorization transformer," in *International Conference on Learning Representations*, 2021, pp. 1–24.

[111] X. Ji, B. Jiang, D. Luo, G. Tao, W. Chu, Z. Xie, C. Wang, and Y. Tai, "ColorFormer: Image colorization via color memory assisted hybrid-attention transformer," in *European Conference on Computer Vision*, 2022, pp. 20–36.

[112] S. Weng, J. Sun, Y. Li, S. Li, and B. Shi, "CT^2: Colorization transformer via color tokens," in *European Conference on Computer Vision*, 2022, pp. 1–16.

[113] X. Kang, T. Yang, W. Ouyang, P. Ren, L. Li, and X. Xie, "DDColor: Towards photo-realistic image colorization via dual decoders," in *IEEE International Conference on Computer Vision*, 2023, pp. 328–338.

[114] V. Santhanam, V. I. Morariu, and L. S. Davis, "Generalized deep image to image regression," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5395–5405.

[115] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5967–5976.

[116] Y. Cao, Z. Zhou, W. Zhang, and Y. Yu, "Unsupervised diverse colorization via generative adversarial networks," in *Machine Learning and Knowledge Discovery in Databases*, 2017, pp. 151–166.

[117] S. Yoo, H. Bahng, S. Chung, J. Lee, J. Chang, and J. Choo, "Coloring with limited data: Few-shot colorization via memory augmented networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 11 275–11 284.

[118] R. Alghofaili, M. Fisher, R. Zhang, M. Lukáč, and L.-F. Yu, "Exploring sketch-based character design guided by automatic colorization," in *Graphics Interface*, 2021, pp. 56–67.

[119] Y. Zhao, L.-M. Po, K.-W. Cheung, W.-Y. Yu, and Y. A. U. Rehman, "SCGAN: Saliency map-guided colorization with generative adversarial network," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 8, pp. 3062–3077, 2021.

[120] X. Jin, Z. Li, K. Liu, D. Zou, X. Li, X. Zhu, Z. Zhou, Q. Sun, and Q. Liu, "Focusing on Persons: Colorizing old images learning from modern historical movies," in *ACM International Conference on Multimedia*, 2021, p. 1176–1184.

[121] G. Kim, K. Kang, S. Kim, H. Lee, S. Kim, J. Kim, S.-H. Baek, and S. Cho, "BigColor: Colorization using a generative color prior for natural images," in *European Conference on Computer Vision*, 2022, pp. 350–366.

[122] H. Lee, D. Kim, D. Lee, J. Kim, and J. Lee, "Bridging the domain gap towards generalization in automatic colorization," in *European Conference on Computer Vision*, 2022, pp. 527–543.

[123] Y. Wang, M. Xia, L. Qi, J. Shao, and Y. Qiao, "PalGAN: Image colorization with palette generative adversarial networks," in *European Conference on Computer Vision*, 2022, pp. 271–288.

[124] C. Saharia, W. Chan, H. Chang, C. Lee, J. Ho, T. Salimans, D. Fleet, and M. Norouzi, "Palette: Image-to-image diffusion models," in *ACM SIGGRAPH*, 2022, pp. 15:1–15:10.

[125] S. Iizuka, E. Simo-Serra, and H. 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*, vol. 35, no. 4, pp. 110:1–110:11, 2016.

[126] G. Özbulak, "Image colorization by capsule networks," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 2150–2158.

[127] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic routing between capsules," in *Advances in Neural Information Processing Systems*, vol. 30, 2017, pp. 3856–3866.

[128] P. Vitoria, L. Raad, and C. Ballester, "ChromaGAN: Adversarial picture colorization with semantic class distribution," in *IEEE Winter Conference on Applications of Computer Vision*, 2020, pp. 2434–2443.

[129] Y. Jin, B. Sheng, P. Li, and C. L. P. Chen, "Broad colorization," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 6, pp. 2330–2343, 2021.

[130] J. Zhao, L. Liu, C. G. M. Snoek, J. Han, and L. Shao, "Pixel-level semantics guided image colorization," in *British Machine Vision Conference*, 2018, pp. 1–12.

[131] J. Zhao, J. Han, L. Shao, and C. G. M. Snoek, "Pixelated semantic colorization," *International Journal of Computer Vision*, vol. 128, pp. 818–834, 2020.

[132] X. Duan, Y. Cao, R. Zhang, X. Wang, and P. Li, "Shadow-aware image colorization," *The Visual Computer*, vol. 40, no. 7, pp. 4969–4979, 2024.

[133] J.-W. Su, H.-K. Chu, and J.-B. Huang, "Instance-aware image colorization," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 7965–7974.

[134] R. Pucci, C. Micheloni, and N. Martinel, "Collaborative image and object level features for image colourisation," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2021, pp. 2160–2169.

[135] M. Xia, W. Hu, T.-T. Wong, and J. Wang, "Disentangled image colorization via global anchors," *ACM Transactions on Graphics*, vol. 41, no. 6, pp. 204:1–204:13, 2022.

[136] X. Dong, W. Li, X. Wang, and Y. Wang, "Learning a deep convolutional network for colorization in monochrome-color dual-

lens system," in *AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 8255–8262.

[137] X. Dong, W. Li, X. Hu, X. Wang, and Y. Wang, "A colorization framework for monochrome-color dual-lens systems using a deep convolutional network," *IEEE Transactions on Visualization and Computer Graphics*, vol. 28, no. 3, pp. 1469–1485, 2020.

[138] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Deep image prior," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 9446–9454.

[139] X. Pan, X. Zhan, B. Dai, D. Lin, C. C. Loy, and P. Luo, "Exploiting deep generative prior for versatile image restoration and manipulation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 11, pp. 7474–7489, 2021.

[140] J. Zhang, C. Xu, J. Li, Y. Han, Y. Wang, Y. Tai, and Y. Liu, "SCSNet: an efficient paradigm for learning simultaneously image colorization and super-resolution," in *AAAI Conference on Artificial Intelligence*, vol. 36, no. 3, 2022, pp. 3271–3279.

[141] M. El Helou and S. Süsstrunk, "BIGPrior: Towards decoupling learned prior hallucination and data fidelity in image restoration," *IEEE Transactions on Image Processing*, vol. 31, pp. 1628–1640, 2022.

[142] Y. Wang, J. Yu, and J. Zhang, "Zero-shot image restoration using denoising diffusion null-space model," in *International Conference on Learning Representations*, 2023, pp. 1–31.

[143] K. C. Chan, X. Xu, X. Wang, J. Gu, and C. C. Loy, "GLEAN: Generative latent bank for image super-resolution and beyond," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 3, pp. 3154–3168, 2023.

[144] L. Yatziv and G. Sapiro, "Fast image and video colorization using chrominance blending," *IEEE Transactions on Image Processing*, vol. 15, no. 5, pp. 1120–1129, 2006.

[145] V. G. Jacob and S. Gupta, "Colorization of grayscale images and videos using a semiautomatic approach," in *IEEE International Conference on Image Processing*, 2009, pp. 1653–1656.

[146] B. Sheng, H. Sun, M. Magnor, and P. Li, "Video colorization using parallel optimization in feature space," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 24, no. 3, pp. 407–417, 2014.

[147] N. Bonneel, J. Tompkin, K. Sunkavalli, D. Sun, S. Paris, and H. Pfister, "Blind video temporal consistency," *ACM Transactions on Graphics*, vol. 34, no. 6, pp. 196:1–196:9, 2015.

[148] C. Lei, Y. Xing, and Q. Chen, "Blind video temporal consistency via deep video prior," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 1083–1093.

[149] S. Xia, J. Liu, Y. Fang, W. Yang, and Z. Guo, "Robust and automatic video colorization via multiframe reordering refinement," in *IEEE International Conference on Image Processing*, 2016, pp. 4017–4021.

[150] Y. Liu, H. Zhao, K. C. Chan, X. Wang, C. C. Loy, Y. Qiao, and C. Dong, "Temporally consistent video colorization with deep feature propagation and self-regularization learning," *Computational Visual Media*, vol. 10, no. 2, pp. 375–395, 2024.

[151] H. Liu, M. Xie, J. Xing, C. Li, and T.-T. Wong, "Video colorization with pre-trained text-to-image diffusion models," *arXiv preprint arXiv:2306.01732*, pp. 1–13, 2023.

[152] J. Li, H. Zhao, Y. Wang, and J. Lin, "Towards photorealistic video colorization via gated color-guided image diffusion models," in *ACM International Conference on Multimedia*, 2024, p. 10891–10900.

[153] N. Ben-Zrihem and L. Zelnik-Manor, "Approximate nearest neighbor fields in video," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 5233–5242.

[154] W.-S. Lai, J.-B. Huang, O. Wang, E. Shechtman, E. Yumer, and M.-H. Yang, "Learning blind video temporal consistency," in *European Conference on Computer Vision*, 2018, pp. 179–195.

[155] S. Liu, G. Zhong, S. De Mello, J. Gu, V. Jampani, M.-H. Yang, and J. Kautz, "Switchable temporal propagation network," in *European Conference on Computer Vision*, 2018, pp. 89–104.

[156] C. Vondrick, A. Shrivastava, A. Fathi, S. Guadarrama, and K. Murphy, "Tracking emerges by colorizing videos," in *European Conference on Computer Vision*, 2018, pp. 402–419.

[157] C. Lei and Q. Chen, "Fully automatic video colorization with self-regularization and diversity," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3748–3756.

[158] E. Casey, V. Pérez, and Z. Li, "The animation transformer: Visual correspondence via segment matching," in *IEEE International Conference on Computer Vision*, 2021, pp. 11 303–11 312.

[159] M. Shi, J.-Q. Zhang, S.-Y. Chen, L. Gao, Y.-K. Lai, and F.-L. Zhang, "Reference-based deep line art video colorization," *IEEE Transactions on Visualization and Computer Graphics*, vol. 29, no. 6, pp. 2965–2979, 2023.

[160] Y. Zhao, L.-M. Po, K. Liu, X. Wang, W.-Y. Yu, P. Xian, Y. Zhang, and M. Liu, "SVCNet: scribble-based video colorization network with temporal aggregation," *IEEE Transactions on Image Processing*, vol. 32, pp. 4443–4458, 2023.

[161] V. Jampani, R. Gadde, and P. V. Gehler, "Video propagation networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3154–3164.

[162] S. Meyer, V. Cornillère, A. Djelouah, C. Schroers, and M. Gross, "Deep video color propagation," in *British Machine Vision Conference*, 2018, pp. 1–14.

[163] B. Zhang, M. He, J. Liao, P. V. Sander, L. Yuan, A. Bermak, and D. Chen, "Deep exemplar-based video colorization," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8044–8053.

[164] Q. Zhang, B. Wang, W. Wen, H. Li, and J. Liu, "Line art correlation matching feature transfer network for automatic animation colorization," in *IEEE Winter Conference on Applications of Computer Vision*, 2021, pp. 3871–3880.

[165] Y. Yang, Z. Peng, X. Du, Z. Tao, J. Tang, and J. Pan, "BiSTNet: semantic image prior guided bidirectional temporal feature fusion for deep exemplar-based video colorization," *arXiv preprint arXiv:2212.02268*, pp. 1–9, 2022.

[166] X. Fang, L. Dai, and J. Tang, "OmniFusion: exemplar-based video colorization using omnimotion and diffusion priors," in *Asian Conference on Computer Vision*, 2024, pp. 1215–1232.

[167] S. Paul, S. Bhattacharya, and S. Gupta, "Spatiotemporal colorization of video using 3d steerable pyramids," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 8, pp. 1605–1619, 2016.

[168] S. Chen, X. Li, X. Zhang, M. Wang, Y. Zhang, J. Han, and Y. Zhang, "Exemplar-based video colorization with long-term spatiotemporal dependency," *Knowledge-Based Systems*, vol. 284, pp. 1–12, 2024.

[169] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," *Master's thesis, Department of Computer Science, University of Toronto*, 2009.

[170] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and F.-F. Li, "ImageNet: a large-scale hierarchical image database," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[171] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International Journal of Computer Vision*, vol. 88, pp. 303–338, 2010.

[172] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: common objects in context," in *European Conference on Computer Vision*, 2014, pp. 740–755.

[173] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," in *Advances in Neural Information Processing Systems*, vol. 27, 2014, pp. 487–495.

[174] F. Yu, A. Seff, Y. Zhang, S. Song, T. Funkhouser, and J. Xiao, "LSUN: construction of a large-scale image dataset using deep learning with humans in the loop," *arXiv preprint arXiv:1506.03365*, pp. 1–9, 2015.

[175] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, "Scene parsing through ade20k dataset," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5122–5130.

[176] V. Bychkovsky, S. Paris, E. Chan, and F. Durand, "Learning photographic global tonal adjustment with a database of input / output image pairs," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 97–104.

[177] H. Chang, O. Fried, Y. Liu, S. DiVerdi, and A. Finkelstein, "Palette-based photo recoloring," *ACM Transactions on Graphics*, vol. 34, no. 4, pp. 139:1–139:11, 2015.

[178] C. Zou, Q. Yu, R. Du, H. Mo, Y.-Z. Song, T. Xiang, C. Gao, B. Chen, and H. Zhang, "SketchyScene: richly-annotated scene sketches," in *European Conference on Computer Vision*, 2018, pp. 438–454.

[179] C. Zou, H. Mo, C. Gao, R. Du, and H. Fu, "Language-based colorization of scene sketches," *ACM Transactions on Graphics*, vol. 38, no. 6, pp. 233:1–233:16, 2019.

[180] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung, "A benchmark dataset and evaluation

methodology for video object segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 724–732.
[181] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev *et al.*, "The kinetics human action video dataset," *arXiv preprint arXiv:1705.06950*, pp. 1–22, 2017.
[182] "Free stock video footage," 2024. [Online]. Available: https://www.videvo.net/
[183] Y. Matsui, K. Ito, Y. Aramaki, A. Fujimoto, T. Ogawa, T. Yamasaki, and K. Aizawa, "Sketch-based manga retrieval using manga109 dataset," *Multimedia Tools and Applications*, vol. 76, pp. 21 811–21 838, 2017.
[184] Y. Wang, "Danbooru 2018 anime character recognition dataset," 2018. [Online]. Available: https://github.com/grapeot/Danbooru2018AnimeCharacterRecognitionDataset
[185] T. Kim, "Anime sketch colorization pair," 2019. [Online]. Available: https://www.kaggle.com/datasets/ktaebum/anime-sketch-colorization-pair
[186] lllyasviel, "sketchkeras," 2017. [Online]. Available: https://github.com/lllyasviel/sketchKeras/
[187] Mukosame, "Anime2sketch," 2021. [Online]. Available: https://github.com/Mukosame/Anime2Sketch/
[188] J. Wang, K. C. Chan, and C. C. Loy, "Exploring clip for assessing the look and feel of images," in *AAAI Conference on Artificial Intelligence*, vol. 37, no. 2, 2023, pp. 2555–2563.
[189] "Laion-aesthtics," 2022. [Online]. Available: https://laion.ai/blog/laion-aesthetics/
[190] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.
[191] Y. Deng, C. C. Loy, and X. Tang, "Image Aesthetic Assessment: an experimental survey," *IEEE Signal Processing Magazine*, vol. 34, no. 4, pp. 80–106, 2017.
[192] N. Zhao, Q. Zheng, J. Liao, Y. Cao, H. Pfister, and R. W. H. Lau, "Selective region-based photo color adjustment for graphic designs," *ACM Transactions on Graphics*, vol. 40, no. 2, pp. 17:1–17:17, 2021.
[193] Z. Ke, Y. Liu, L. Zhu, N. Zhao, and R. W. Lau, "Neural preset for color style transfer," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2023, pp. 14 173–14 182.
[194] jamriska, "ebsynth," 2018. [Online]. Available: https://github.com/jamriska/ebsynth/
[195] S. Yang, Y. Zhou, Z. Liu, and C. C. Loy, "Rerender A Video: zero-shot text-guided video-to-video translation," in *ACM SIGGRAPH Asia*, 2023, pp. 95:1–95:11.
[196] C. Qi, X. Cun, Y. Zhang, C. Lei, X. Wang, Y. Shan, and Q. Chen, "Fatezero: Fusing attentions for zero-shot text-based video editing," in *IEEE International Conference on Computer Vision*, 2023, pp. 15 886–15 896.

Yu Cao received the B.Eng. degree in communication engineering from the Qingdao Institute of Technology, Qingdao, China, in 2017, and the M.Eng. degree in communication and information system from the Xidian University, Xi'an, China, in 2020. He is currently a Ph.D. candidate in the School of Fashion and Textiles at The Hong Kong Polytechnic University, Hong Kong. His research interests include computer graphics, computer vision, and deep learning.

Xin Duan received the B.Eng. degree in software engineering from Harbin Institute of Technology, China, in 2019, and the M.Sc. degree in computer science from the University of Hong Kong, Hong Kong in 2020. She is currently pursuing the Ph.D. in computing with The Hong Kong Polytechnic University, Hong Kong. Her current research interests include video object segmentation, diffusion models, and image colorization.

Xiangqiao Meng received the B.Eng. degree in coastal engineering and the M.Sc. degree in aerospace information technology both from the Zhejiang University, Hangzhou, China, in 2019 and 2022, respectively. He is currently pursuing the Ph.D. degree in computing with The Hong Kong Polytechnic University, Hong Kong. His current research interests include image colorization, diffusion models, and image synthesis.

P. Y. Mok (Member, IEEE) received the B.Eng. degree (Hons.) and the Ph.D. degrees in industrial and manufacturing systems engineering from The University of Hong Kong, in 1998 and 2002, respectively. She is currently an Associate Professor with The Hong Kong Polytechnic University, Hong Kong. Her current research interests include fashion 2D and 3D CAD, digital human modeling, cloth simulation, deep learning, sketch and pattern designs, computer graphics in fashion, and fashion design and synthesis.

Ping Li (Member, IEEE) received the Ph.D. degree in computer science and engineering from The Chinese University of Hong Kong, Hong Kong, in 2013. He is currently an Assistant Professor with the Department of Computing and an Assistant Professor with the School of Design, The Hong Kong Polytechnic University, Hong Kong. He has published over 250 top-tier scholarly research articles (e.g., TVCG, TPAMI, TIP, TNNLS, TMI, TMM, TCSVT, TCYB, TBME, TSMC, TII, AAAI, CVPR, ICCV, ECCV, NeurIPS, Nature Metabolism), pioneered several new research directions, and made a series of landmark contributions in his areas. He has an excellent research project reported by the *ACM TechNews*, which only reports the top breakthrough news in computer science worldwide. More importantly, however, many of his research outcomes have strong impacts to research fields, addressing societal needs and contributed tremendously to the people concerned. His current research interests include image/video stylization, colorization, artistic rendering and synthesis, computational art, and creative media.

Tong-Yee Lee (Senior Member, IEEE) received the Ph.D. degree in computer engineering from Washington State University, Pullman, in 1995. He is currently a Chair Professor with the Department of Computer Science and Information Engineering, National Cheng-Kung University (NCKU), Tainan, Taiwan. He leads the Computer Graphics Group, Visual System Laboratory, NCKU (http://graphics.csie.ncku.edu.tw). His current research interests include computer graphics, non-photorealistic rendering, medical visualization, virtual reality, and media resizing. He is a Senior Member of the IEEE and a Member of the ACM. He is an Associate Editor of the *IEEE Transactions on Visualization and Computer Graphics*.