The Visual Computer (2024) 40:4969–4979
https://doi.org/10.1007/s00371-024-03500-5

RESEARCH

Shadow-aware image colorization

Xin Duan[1] · Yu Cao[2] · Renjie Zhang[1] · Xin Wang[1] · Ping Li[1,3]

Accepted: 16 May 2024 / Published online: 4 June 2024

Abstract
Significant advancements have been made in colorization in recent years, especially with the introduction of deep learning technology. However, challenges remain in accurately colorizing images under certain lighting conditions, such as shadow. Shadows often cause distortions and inaccuracies in object recognition and visual data interpretation, impacting the reliability and effectiveness of colorization techniques. These problems often lead to unsaturated colors in shadowed images and incorrect colorization of shadows as objects. Our research proposes the first shadow-aware image colorization method, addressing two key challenges that previous studies have overlooked: integrating shadow information with general semantic understanding and preserving saturated colors while accurately colorizing shadow areas. To tackle these challenges, we develop a dual-branch shadow-aware colorization network. Additionally, we introduce our shadow-aware block, an innovative mechanism that seamlessly integrates shadow-specific information into the colorization process, distinguishing between shadow and non-shadow areas. This research significantly improves the accuracy and realism of image colorization, particularly in shadow scenarios, thereby enhancing the practical application of colorization in real-world scenarios.

Keywords Colorization · Shadow detection · Transformer

1 Introduction

Enhancing grayscale images with color components is the key to unlocking richer semantics and achieving visually striking appearances. Over the past few decades, the pursuit of effective image colorization has been a longstanding research problem in computer graphics. Early endeavors involved handcrafted features [8, 19, 27, 36, 46], evolving toward high-level semantic features [26, 29, 41, 49, 57] to produce vivid and diverse colorization results. In the current era dominated by deep learning, the emphasis on image semantics has led to the incorporation of specifically designed semantic-aware structures [29, 41, 49] and the utilization of large-scale dataset pre-trained networks as backbone models [16, 20, 40] for extracting more semantic features.

The current success of colorization research primarily focuses on common targets such as “blue sky" and “green grass" achieved through large-scale datasets like ImageNet [13] or COCO-Stuff [5], with backbone networks implicitly containing semantic awareness. Although existing learning-based methods [12, 26, 55] have achieved good results in the colorization task, none of the existing work considered rare illumination cases, particularly "shadow", which is not adequately recognized or addressed by current techniques. This neglect of shadow leads to distortions and inaccuracies in object awareness, especially in real-world scenarios with shaded images. Consequently, there is a clear need to expand the research to encompass the full spectrum of possible colorization subjects, including those with rare illumination conditions, to improve the overall effectiveness and accuracy of colorization methods.

✉ Ping Li
p.li@polyu.edu.hk

Xin Duan
hizuka.duan@connect.polyu.hk

Yu Cao
yu-daniel.cao@connect.polyu.hk

Renjie Zhang
renjie.zhang@connect.polyu.hk

Xin Wang
xin1025.wang@connect.polyu.hk

[1] Department of Computing, The Hong Kong Polytechnic University, Kowloon, Hong Kong

[2] School of Fashion and Textiles, The Hong Kong Polytechnic University, Kowloon, Hong Kong

[3] School of Design, The Hong Kong Polytechnic University, Kowloon, Hong Kong

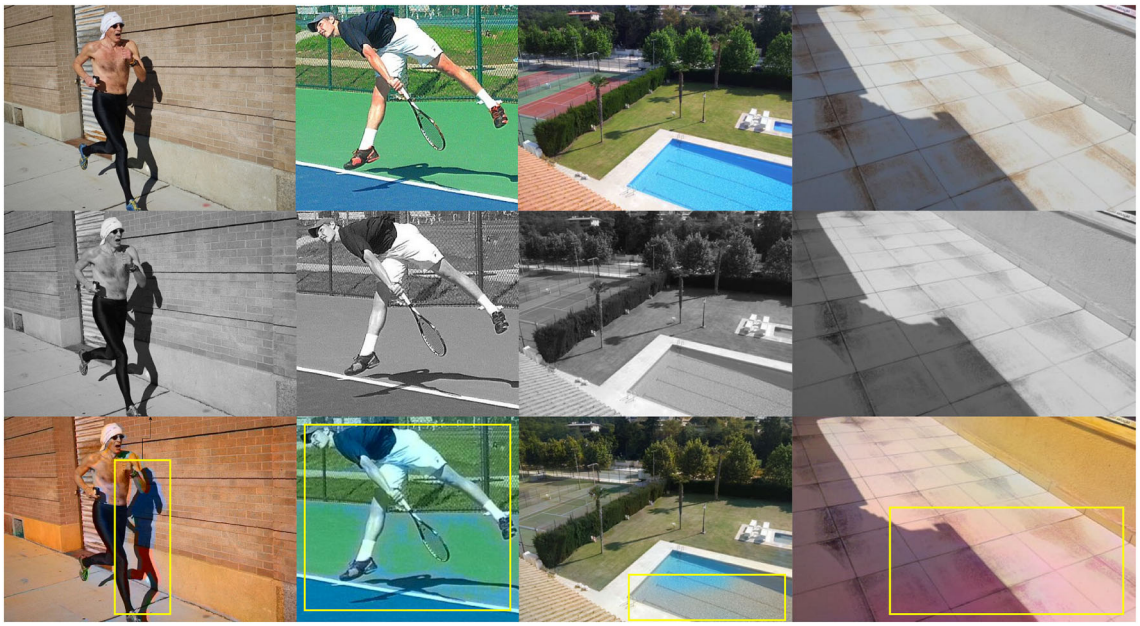


Fig. 1 The image above displays the ground-truth image (first row), input image (second row), and the output of previous methods (third row), including BigColor [31], CT^2 [47], ColorFormer [28], and DDColor [29]. The colorization process has led to the misinterpretation of shadow areas as objects, resulting in them being colored with unusual red and blue hues (column 1). Furthermore, directly fine-tuning existing trained models on shadow images leads to the overall image color appearing desaturated in the presence of shadows (columns 2 and 3), and in images with a significant proportion of shadows, incorrect color hues have been injected (column 4)

Detecting shadows faces significant challenges due to their lack of specific shapes, colors, or textures and their intensity being slightly lower than the surrounding areas. Many low-level vision researchers address shadow detection in separate research from common object detection, using intricate network architectures [11, 23, 43, 58] trained directly on shadow datasets [22, 24, 44, 59] with labeled shadow regions. Despite these advancements, the colorization community has historically treated shadow as a separate and neglected task, with few works considering shadow in colorization. However, shadows are important in semantic preservation in light transmission laws or geometrically precise object structures.

In this paper, we introduce the novel concept of shadow-aware colorization. Existing colorization methods exhibit surprising inadequacies and instability when confronted with shadowed scenes in real-world colorization tasks, leading to semantic understanding impediments and distortions. Conversely, directly fine-tuning an existing colorization model with shadowed images can easily cause the model to produce unsaturated images (see columns 2 and 3 of Fig. 1). We identify two key issues in shadow-aware colorization. The first issue is how to incorporate shadow information with existing general semantic understanding, and the second is how to preserve saturation and colorfulness, as shadow often appears colorless.

We aim to address the aforementioned issues by proposing a novel end-to-end framework for shadow-aware colorization. Our key insight is that a clear separation of shadow can significantly enhance colorization performance. Unlike existing methods that learn to colorize the entire image, we design two protocols within our *shadow-aware block* to colorize areas with and without shadows, respectively. Non-shadow areas are colorized using color features to achieve saturation and diversity. In contrast, shadow areas are processed separately to ensure that colorless shadow features do not hinder the overall saturation. This separation design offers two main advantages. First, colorizing non-shadow areas is substantially easier as it does not need to handle complex illumination distortion. Second, using shadow separation as input allows the network to produce colorful results even with the dominance of low-saturation shadow samples.

Our main contributions are summarized as follows:

- We introduce the concept of “Shadow-Aware Colorization" to address the limitations of existing colorization methods when dealing with shadowed scenes, particularly in real-world scenarios with shaded images.
- We design a novel end-to-end framework with dual branches: one for outputting the binary shadow mask and the other for shadow-aware colorization, which considers shadow features to predict missing color channels.

- We propose a novel *shadow-aware block* with shadow separation to enhance colorization performance by preserving saturation and colorfulness, offering significant advantages in handling complex illumination distortion and producing colorful results even with low-saturation shadow samples.

2 Related work

Conditional colorization Conditional colorization is a kind of image generation or editing task under various kinds of user controls including reference images [1, 6, 7, 17, 21, 50], color strokes (or scribbles) [34, 39, 42, 51, 54, 56], and texts [4, 9, 60]. It has evolved from the traditional optimization-based approach to the current learning-based approach and is widely applied in image and line art colorization. Some research focuses on implementing specific controls to improve colorization quality. For instance, Kim et al. [30] introduced an edge-enhancing network to reduce color-bleeding artifacts. Additionally, Huang et al. [25] proposed a unified framework capable of managing various controls, including stroke, exemplar, text, and combinations. However, no research currently considers shadows in the colorization task.

Automatic colorization Colorizing grayscale images is challenging since objects in these images can have a wide range of colors in reality. Determining how to assign realistic and appealing colors is a difficult problem. However, with the availability of large-scale datasets and advancements in deep neural networks (DNNs), it has become feasible to colorize grayscale images in a data-driven manner. These unconditional models, as described in the works of [12, 15, 33, 55], were capable of predicting the most plausible colors for the given input image without any laborious color annotations provided by a user. To address the issue of results being dominated by frequently occurring colors, researchers have turned to the use of generative models to design their methods [14, 31, 38, 45, 48] which can learn the inherent color distribution of images, thereby enabling models to produce a wider range of diverse colorization results.

Semantic understanding in colorization Image colorization research is advancing rapidly, primarily due to introducing a high-performance visual backbone network and using semantic information to inspire upstream vision tasks [28]. A strong ability to understand semantics could potentially enhance a model's colorization performance. Therefore, existing colorization networks often incorporate semantic understanding modules to process input images. Existing works designed for image colorization can be categorized into three levels, including global level [26, 29, 49], pixel level [57], and instance level [41]. In this study, we integrate shadow semantic understanding into the colorization pipeline to enhance the colorization outcomes.

Shadow construction and shadow types Shadows are common when a surface is not directly illuminated by light, and they have garnered significant attention in the research field. Shadows can be broadly categorized into self-shadow and cast shadow, as described by Hu et al. [24]. Self-shadow refers to an object blocking its light, while cast shadow occurs when one object casts shadows onto another surface. Cast shadow can be further classified into the umbra and penumbra components [2, 3], with the umbra representing a region where no light can directly reach from the light source, and the penumbra being a smooth transition from no light to full light. In real-world scenarios, penumbra regions are typically found in shadow boundaries and have minimal impact on the overall shadow appearance. In the realm of deep learning-based shadow detection and removal approaches [11, 22, 23, 35, 43, 44, 58, 59], the distinction between umbra and penumbra was often discarded, with both components being collectively referred to as cast shadow. This general protocol is also followed in our current work, where the distinction between the two is not made.

Shadow detection The significance of shadow detection in low-level vision has been widely acknowledged in previous research, as shadows can greatly influence the perceived colors in an image. Several studies have concentrated on developing specialized networks [11, 23, 43, 58] to detect or eliminate shadows from images accurately. Despite the extensive research on shadow detection, shadows are currently not considered in the context of colorization. Addressing the impact of shadows on the colorization process is essential for producing accurate and visually appealing results, ensuring that shadows do not distort the colors applied to the image. By integrating a shadow-aware block into the colorization process, our proposed model aims to separately consider the impact of shadows on the final colorized images.

3 Method

3.1 Overview

Our network operates by taking a grayscale image $I \in \mathbb{R}^{H\times W\times 1}$ as input and then proceeds to predict its binary shadow mask $Y_s \in \mathbb{R}^{H\times W\times 1}$ and two missing color channels $Y_c \in \mathbb{R}^{H\times W\times 2}$ in the CIELAB color space, all in an end-to-end fashion. The network architecture, as depicted in Fig. 2, consists of dual branches: a *shadow detection branch* responsible for outputting the binary shadow mask, and a

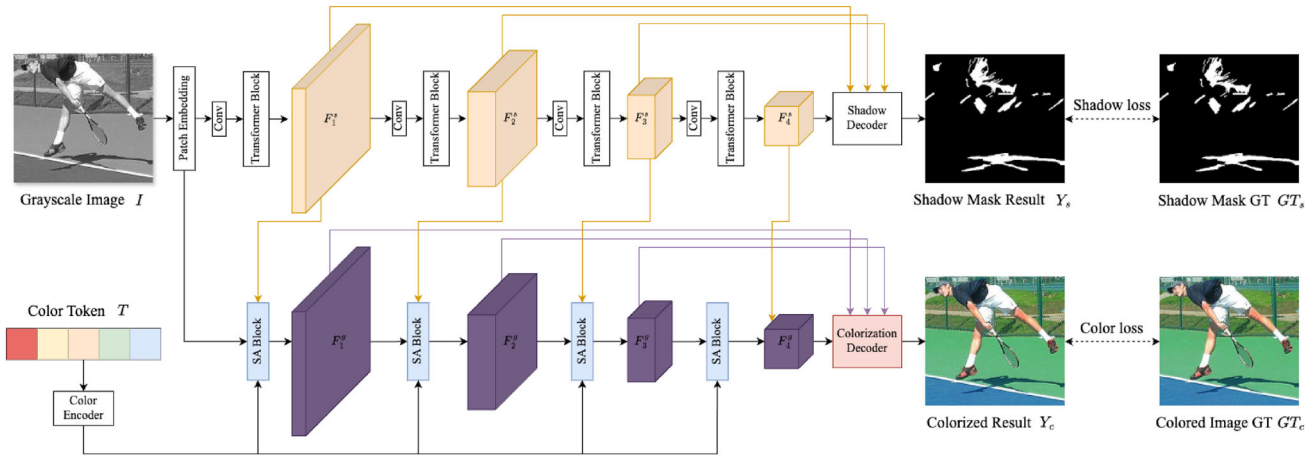


Fig. 2 The structure of the dual-branch shadow-aware colorization network involves two encoders that are not entirely separate, as they incorporate feature transfer in the middle to achieve shadow-aware colorization. This feature transfer mechanism allows for sharing certain extracted features between the two encoders, enabling the model to effectively perform shadow-aware colorization by leveraging relevant information from both tasks. The interconnected nature of the encoders facilitates the extraction of task-specific features while also enabling the integration of shared features to enhance the colorization process with shadow awareness

shadow-aware colorization branch designed to predict the missing color channels while considering shadow features.

3.2 Dual-branch shadow-aware colorization network

Given that features required for colorization and shadow detection tasks differ to some extent, with colorization needing to accurately recover true colors and shadow detection requiring binary activation only on shadow parts, we apply dual encoders to extract task-specific features in the encoding part combined with two light-weight task-specific decoders for each output.

Shadow detection branch We build a transformer-based encoder that outputs multi-level features in the encoding part. The input image I is sequenced using patch embedding. We use a 3×3 convolution operation in each layer to reduce spatial resolution, followed by a transformer block to extract shadow features. The shadow feature at level i is represented as F_i^s, having a resolution of $\frac{H}{2^{i+1}} \times \frac{W}{2^{i+1}} \times C_i$, where i varies from 1 to 4, and C_{i+1} exceeds C_i. The Mixformer is the backbone, as outlined in [52]. In the decoding phase, multi-level features can amplify semantic information by merging low-resolution detailed features with high-resolution coarse features. This is accomplished through a sequence of stacked upsampling layers that reinstate the spatial resolution of the image features. Each upsampling layer comes with a shortcut connection to the corresponding stage of the encoder. The shadow decoder's output is denoted by the binary shadow mask $Y_s \in \mathbb{R}^{H \times W \times 1}$, which retains the same spatial resolution as the input image.

Shadow-aware colorization branch Following the patch embedding process, the grayscale image feature F_i^g is derived, integrating the shadow feature F_i^s alongside the color token T into our proposed shadow-aware block (Section 3.3). We construct color token T following [47] and set the same size as F_i^g. This integration is strategically designed to facilitate the extraction of shadow-aware color features, ensuring that the colorization process is attuned to the presence of shadows in the input image. The colorization decoder shares the same structure as the shadow decoder, but their weights are not shared. The model can effectively produce task-specific outputs by maintaining separate weights for the colorization and shadow detection decoders, optimizing the accuracy and precision of both colorization and shadow detection processes.

3.3 Shadow-aware block

Here, we explore the fusion of shadow features with image features to enhance colorization. Our primary insight is that the distinct separation of shadows can significantly improve colorization performance. In Fig. 3, we illustrate the details of our shadow-aware block. This block is utilized at multiple layers of the colorization branch, but we only present it at the i layer for simplicity. Applying this module to all other layers is straightforward.

The shadow-aware block requires three inputs: the grayscale image feature F_i^g, a color token T, and the shadow feature F_{i+1}^s. With shadow awareness, it then produces the grayscale image feature F_{i+1}^g at the next level. A predefined threshold equal to 0.5 is established to calibrate the shadow feature. This feature undergoes softmax normalization before the

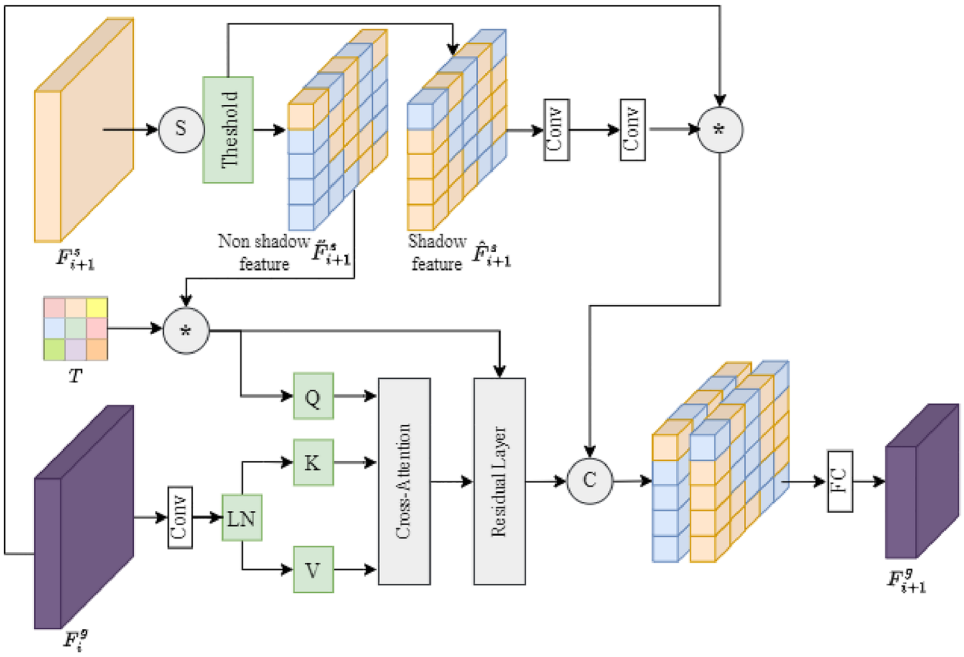


Fig. 3 Shadow-aware block. This block needs three inputs: the grayscale image feature F_i^g, a color token T, and the shadow feature F_{i+1}^s. It subsequently generates the grayscale image feature F_{i+1}^g at the succeeding level, which is shadow-aware

threshold function separates F_{i+1}^s into the shadow feature $\hat{F}_{i+1}^s$ and the non-shadow feature $\tilde{F}_{i+1}^s$. The shadow feature $\hat{F}_{i+1}^s$ is obtained by setting feature values to 0 where the original feature value is below the threshold. Conversely, the non-shadow feature $\tilde{F}_{i+1}^s$ is obtained by setting feature values to 0 where the pixel value exceeds the threshold.

We employ different protocols to integrate shadow and non-shadow features. The non-shadow feature $\tilde{F}_{i+1}^s$ first undergoes a dot product operation with the color token T to produce query vector Q. This is followed by a cross-attention operation and a residual connection with the grayscale feature F_i^g serving as K and V. The introduction of the color token aims to mitigate undersaturation in non-shadow areas. The process can be formulated by:

$$\tilde{F}_{i+1}^g = \text{softmax}\left(Q \cdot K^T\right) V + (\tilde{F}_{i+1}^s \cdot T). \tag{1}$$

As shadows tend to be colorless, we define a separate protocol for shadow area colorization. The shadow features $\hat{F}_{i+1}^s$ are processed through two convolutional layers and a dot product operation with the grayscale feature F_i^g. The equation is as follows:

$$\hat{F}_{i+1}^g = \text{Conv}(\text{Conv}(\hat{F}_{i+1}^s)) \cdot F_i^g. \tag{2}$$

Finally, the two outputs are concatenated and passed through a fully connected layer for dimension reduction. The resulting output, the F_{i+1}^g grayscale image feature, is now equipped with shadow awareness:

$$F_{i+1}^g = \text{FC}(\text{Concat}((\tilde{F}_{i+1}^g)(\hat{F}_{i+1}^g)). \tag{3}$$

3.4 Loss function

In this section, we delve into the loss functions used in our dual- branch shadow-aware colorization network, which comprises two functions: shadow loss and color loss.

For the optimization of the binary shadow mask, the loss is calculated as follows:

$$\mathcal{L} = \mathcal{L}_{\text{BCE}}(Y_s, \text{GT}_s) + \mathcal{L}_{\text{Hinge}}(Y_s, GT_s), \tag{4}$$

where $\mathcal{L}_{\text{BCE}}(\cdot)$ represents the binary cross-entropy (BCE) loss, and $\mathcal{L}_{\text{Hinge}}(\cdot)$ denotes the Lovász-Hinge loss. These two losses are combined to optimize the shadow mask, where Y_s is the predicted shadow mask and GT_s is the ground-truth shadow mask. For colorization, we follow the approach of Zhang et al. [56] and adopt the smooth-ℓ_1 loss to reduce the likelihood of overly saturated color candidates. The smooth-ℓ_1 loss is defined as:

$$L_\delta = \begin{cases} \frac{1}{2}(y-x)^2 & \text{if } |(y-x)| < \delta \\ \delta((y-x) - \frac{1}{2}\delta) & \text{otherwise,} \end{cases} \tag{5}$$

where x and y represent the predicted and ground-truth color values, respectively. The squared difference is used if the absolute difference is less than δ. A linear function of the difference is used if the absolute difference is greater than or equal to δ. This loss function is less sensitive to outliers, making it suitable for colorization tasks where overly saturated colors are undesirable.

4 Experimental results

4.1 Shadow image collection

To ensure the accuracy and completeness of our model with shadow awareness, we undertook a thorough process of identifying and gathering relevant datasets. We recognized that shadows are a unique labeling type in the current research field and have received less attention from general large benchmark datasets such as ImageNet [13] or COCO-Stuff [5]. As a result, we collected **the first** united shadow dataset for colorization by incorporating data from widely used image shadow detection datasets including CUHK-Shadow [24], ISTD [44], SBU [22], UCF [59], and video shadow detection datasets ViSha [10] and STICT [37]. These datasets were originally intended for image or video shadow detection, with each sample pair consisting of a fully colored image and its corresponding binary shadow mask. For example, CUHK-Shadow [24] is a real-world image shadow dataset with self-shadows. We only adopted the first and last frame for video datasets to prevent an overabundance of similar samples.

Our united dataset comprises 18,111 shadow images, with 4,483 for testing and 13,628 for training, encompassing diverse shadow samples showcasing various items, scenarios, artifacts, and shadows of different scales. This comprehensive selection accurately represents the complexities of the real world and the interactions between shadows and objects. By creating a united dataset, we ensure that our model is based on a robust and comprehensive collection of real-world shadow masks, thereby enhancing the accuracy and completeness of our research. Additionally, the potential impact of having a united dataset on this topic is significant, as it allows for more comprehensive and accurate analysis and development of colorization techniques specifically tailored to address the challenges posed by shadows.

4.2 Implementation details

We constructed our network by using the PyTorch and PyTorch-Lighting frameworks. The parameters of the transformer blocks of the shadow detection branch were initialized using the weights from the MiT-B4 [52]. The remaining parameters (shadow-aware colorization branch and two decoders) were randomly initialized using "Xavier" methods [18]. For optimization, we employed the Adan [53] optimizer with an initial learning rate of 1.5×10^{-5} and a weight decay of 0.02. We set the resolution of input and output images as 512×512. All experiments were trained for 100 epochs on a single NVIDIA GeForce RTX 3090 GPU.

4.3 Evaluation metrics

In line with the experimental setting of existing colorization methods, we utilized a range of evaluation metrics to quantify the quality of our colorization results. These metrics included the peak signal-to-noise ratio (PSNR), Structural Similarity Index Measure (SSIM), Learned Perceptual Image Patch Similarity (LPIPS), Fréchet inception distance (FID), and colorfulness score (CF) following [47].

4.4 Comparison with SOTA methods

Comparative Methods. We evaluated our method by comparing it with four advanced transformer-based colorization methods, namely ColTran [32], CT^2 [47], ColorFormer [28], DDColor [29], and one generative approach, BigColor [31]. To ensure a meaningful comparison, we retrained their models on our united shadow training set and reported results by testing all methods on our united test shadow set.

Quantitative Comparison. As shown in Table 1, our method outperforms the existing colorization methods on our united shadow test dataset across multiple evaluation metrics. Specifically, our method achieves the highest PSNR (23.58), indicating superior fidelity to the ground truth images compared to the comparative methods. Additionally, our method demonstrates the highest SSIM (0.92), indicating a high level of structural similarity with the ground-truth images. Our method achieves the lowest LPIPS score (0.20) regarding perceptual image patch similarity, indicating the closest perceptual similarity. Furthermore, our method outperforms the comparative methods regarding distribution similarity, as indicated by the lowest FID score (4.18). While our method excels in fidelity and perceptual similarity, it also achieves a high colorfulness score (33.75), indicating vivid and visually appealing colorization results. Additionally, our method demonstrates a low colorfulness score variance ($\triangle$ CF) of 1.64, further highlighting the consistency and quality of the colorization output. Our method showcases the best performance across various evaluation metrics, demonstrating its effectiveness in producing high-quality colorized images on the united shadow test dataset.

Qualitative Comparison. In the qualitative visual comparison of Fig. 4, our colorization results exhibit more natural and vivid colors, especially in areas with shadows, compared to the outputs of other methods. Our method captures and enhances the color details in shadowed regions, resulting in more realistic and visually appealing outcomes. For example, our method achieves better results in terms of color consistency, as other methods often fail to consistently colorize images in shadow areas, resulting in localized colormissing.

Table 1 Quantitative comparison between our method and existing colorization methods on our united shadow image dataset

Methods	Our united shadow test dataset					
	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓	CF ↑	△ CF ↓
ColTran [32]	21.46	0.79	0.30	6.57	30.95	4.49
CT^2 [47]	21.88	0.82	0.24	5.73	31.88	3.53
ColorFormer [28]	20.79	0.89	0.23	6.61	30.69	5.58
BigColor [31]	20.38	0.80	0.31	8.60	32.72	2.61
DDColor [29]	22.57	0.91	0.22	6.07	32.13	2.56
⋆ Ours	23.58	0.92	0.20	4.18	33.75	1.64

↑, the larger the value, the better the performance, while ↓, means the opposite

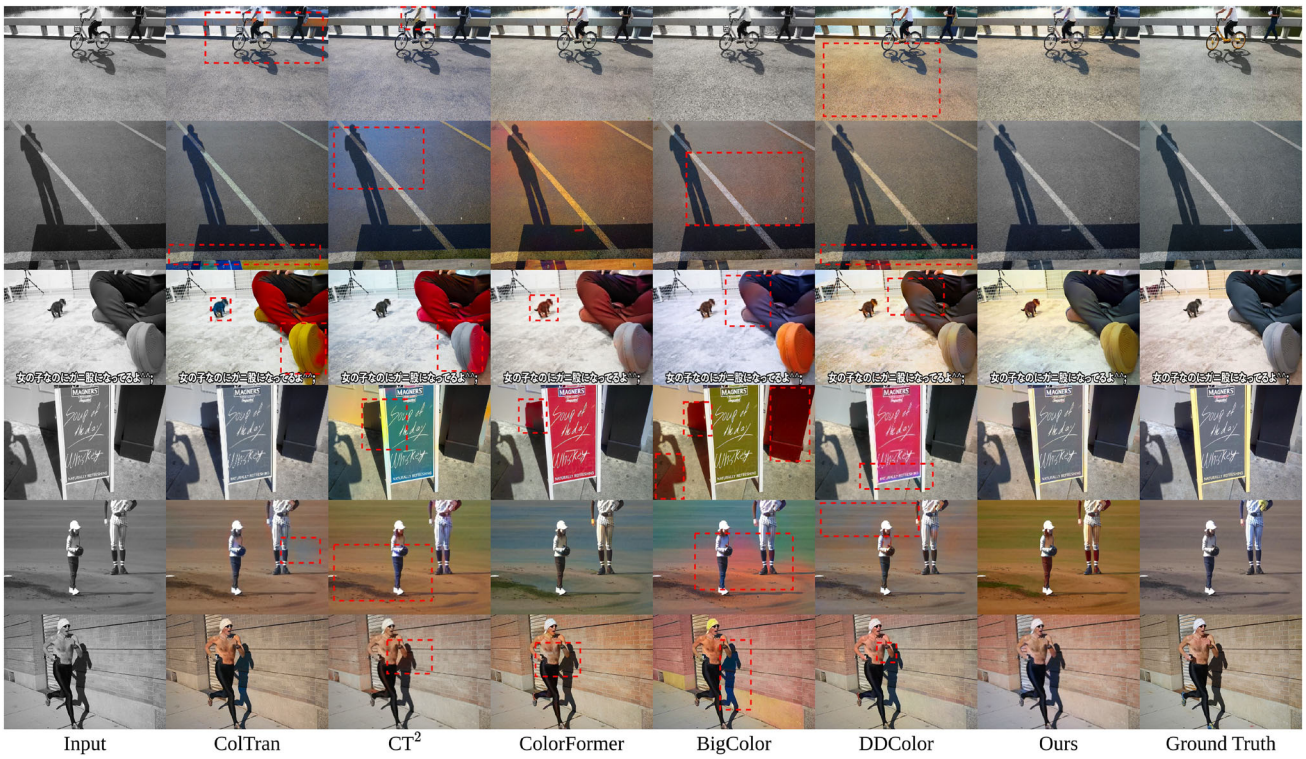


Fig. 4 Qualitative comparison of colorization results generated by our and other methods, including ColTran [32], CT^2 [47], ColorFormer [28], BigColor [31], and DDColor [29]. Our method produces colorized images with more natural and vivid colors, particularly in the presence of shadows, compared to the results obtained from other methods

In contrast, our method maintains better color coherence in these regions. Furthermore, when shadow areas become dominant in the image, some methods mistakenly colorize shadow areas with weird colors, such as the bottom shadow area in the second row, while ours remains naturally preserved. Overall, the qualitative comparison highlights our method's ability to produce colorized images with superior color accuracy, vibrancy, and naturalness, particularly in shadows, setting it apart from other existing methods.

Penumbra Regions. Shadows produced under illumination from multiple area light sources often includepenumbra, an intricate gray-scale shadow pattern that cannot be binary-classified. As this phenomenon frequently appears in everyday situations, we sought to visually validate whether our methods can effectively handle such soft-shadow cases. In Fig. 5, which contains varying degrees of tough penumbra regions, our performance remains unaffected. We believe this is because our binary classification of shadows already provides a sufficient penumbra approximation in the colorization task. This observation also aligns with the shadow detection community [10, 11, 35], which also regards shadow as a simple binary classification problem.

Fig. 5 Additional results of our method showcasing penumbra regions. The first and third rows display the ground-truth colorful images, while the second and fourth rows show our colorization results. The distinctive penumbra regions in each sample are highlighted using a dashed box

4.5 User study

We conducted a user study to investigate the human evaluation for each colorization method. Specifically, we compared our method with ColTran [32], CT^2 [47], ColorFormer [28], BigColor [31], and DDColor [29]. We randomly selected 50 grayscale test images from our collected shadow dataset, and the colored results were displayed to 20 participants. Subjects selected the visually best-colorized image and rated it on a Likert scale of 1–5 in the global color effect and color tone in the shadow region. All results of different methods were randomly shuffled. As shown in Fig. 6, our method is preferred by a wider range of users than the state-of-the-art methods.

4.6 Ablation study

To evaluate the effectiveness of the proposed dual-branch shadow-aware colorization network, we conduct an ablation study to assess the impact of its key components. Specifically, we investigate the shadow detection branch's contributions and the shadow-aware block's integration.

We first train the network solely on the colorization branch by modifying the network structure and removing the input shadow feature F_i^s. The quantitative results are reported in the first row of Table 2 and Fig. 7b. It is observed that while the network can produce colorful results, it mistakenly colors shadow areas with incorrect colors due to the lack of shadow awareness. Next, we introduce the shadow detection branch without the shadow-aware block. This is achieved by replacing our Shadow-Aware Block with a naive cross-attention operation, with the shadow feature F_{i+1}^s dot product with the color feature T as Q, and the grayscale feature F_i^g as K and V. The results are reported in the second row of Table 2 and Fig. 7c. It is observed that the introduction of the shadow feature positively contributes to color correctness; however, the overall colorfulness is negatively impacted. Thus, Fig. 7c appears unsaturated. Finally, we evaluate the complete version of our network with both the shadow detection branch and shadow-aware block, achieving the best performance. It can be observed from Fig. 7d that the overall colorization results appear more natural, with shadow areas correctly colored and non-shadow areas exhibiting vivid and natural colors.

Table 2 Ablation study of our network

SD Branch	SA Block	FID ↓	CF ↑	△ CF ↓
×	×	5.13	31.87	2.31
✓	×	4.56	29.69	3.07
⋆ ✓	⋆ ✓	4.18	33.75	1.64

The first row shows the results of training the network solely on the colorization branch; the second row demonstrates the results of introducing the shadow detection branch without the shadow-aware block. Finally, ⋆ shows the results of the complete network with both the shadow detection branch and shadow-aware block, yielding the best results

4.7 Limitation

As illustrated in Fig. 8, our model still encounters challenges when processing images with intricate or confusing shadow patterns, resulting in inaccurate shadow masks. In the two leftmost samples, the model struggles to colorize certain parts of the shadowed areas accurately. Moreover, our model's semantic comprehension remains inadequate, as in the right-

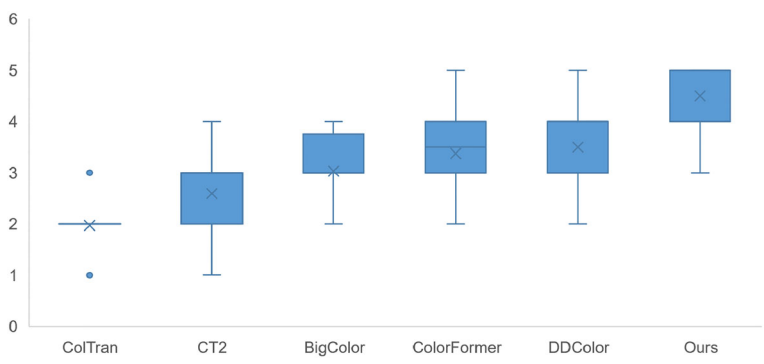


Fig. 6 Overall user study results from 20 participants. Each participant rates on a five-point scale according to the visual assessment of colored results in terms of the global color effect and color tone in the shadow region of six methods including ColTran [32], CT^2 [47], BigColor [31], ColorFormer [28], DDColor [29], and ours. This indicates that our method achieves the best results through human assessment

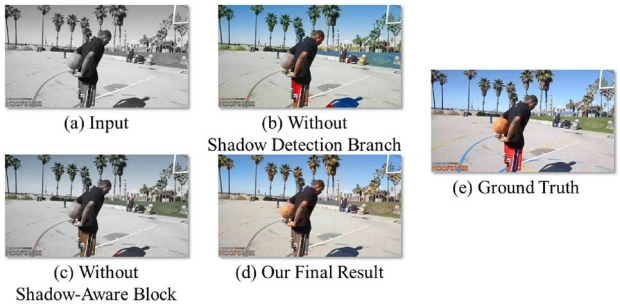


Fig. 7 Visual results of ablation. **a** Input to the network. **b** The results of training the network solely on the colorization branch, which produces colorful but incorrectly colored shadow areas. **c** Introducing the shadow detection branch without the shadow-aware block improves color correctness but negatively affects overall colorfulness, resulting in unsaturated images. **d** Shows the results of the complete network with both the shadow detection branch and shadow-aware block, yielding the most natural colorization results

Fig. 8 Limitations in processing images with intricate or confusing shadow patterns. The left two samples show the model's struggle to accurately colorize certain parts of the shadowed areas, while the rightmost sample demonstrates the model's inadequate semantic comprehension

most sample, where the model erroneously assigns a green color to a shadow area resembling a tree, indicating a failure to distinguish it from an actual tree. To achieve further improvement, additional semantic supervision may be necessary to help the network better comprehend such complex scenarios.

5 Conclusion

This study presents the initial solution to the challenges of image colorization in shadow scenarios, an area that has been largely overlooked in previous research. By integrating shadow-specific knowledge with the general understanding implied by the colorization branch, our proposed model demonstrates significant performance improvements over existing colorization models. Experimentation on our united shadow dataset validates the effectiveness of our proposed dual-branch shadow-aware colorization network and shadow-aware block, highlighting its capability to address the limitations observed in current colorization models. We are confident that the outcomes of our research open up new possibilities for colorization in real-world applications.

Acknowledgements The authors would like to thank the editors and anonymous reviewers for their insightful comments and suggestions. This work was supported by The Hong Kong Polytechnic University under Grants P0048387, P0042740, P0044520, P0043906, P0049586, and P0050657.

Funding Open access funding provided by The Hong Kong Polytechnic University

Declarations

Conflict of interest All authors declare that they have no conflict of interest.

References

1. Akita, K., Morimoto, Y., Tsuruno, R.: Colorization of line drawings with empty pupils. Comput. Graph. Forum **39**(7), 601–610 (2020)
2. Assarsson, U., Akenine-Möller, T.: A geometry-based soft shadow volume algorithm using graphics hardware. ACM Trans. Graph. **22**(3), 511–520 (2003)
3. Baba, M., Mukunoki, M., Asada, N.: Shadow removal from a real image based on shadow density. In: ACM SIGGRAPH Posters, p. 60 (2004)
4. Bahng, H., Yoo, S., Cho, W., Park, D.K., Wu, Z., Ma, X., Choo, J.: Coloring with words: guiding image colorization through text-based palette generation. In: European Conference on Computer Vision, pp. 443–459 (2018)
5. Caesar, H., Uijlings, J., Ferrari, V.: COCO-Stuff: thing and stuff classes in context. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1209–1218 (2018)
6. Cao, Y., Meng, X., Mok, P.Y., Lee, T.Y., Liu, X., Li, P.: AnimeDiffusion: anime diffusion colorization. IEEE Trans. Vis. Comput. Graph. (2024). https://doi.org/10.1109/TVCG.2024.3357568
7. Cao, Y., Tian, H., Mok, P.Y.: Attention-aware anime line drawing colorization. In: IEEE International Conference on Multimedia and Expo, pp. 1637–1642 (2023)
8. Charpiat, G., Hofmann, M., Schölkopf, B.: Automatic image colorization via multimodal predictions. In: European Conference on Computer Vision, pp. 126–139 (2008)
9. Chen, J., Shen, Y., Gao, J., Liu, J., Liu, X.: Language-based image editing with recurrent attentive models. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 8721–8729 (2018)
10. Chen, Z., Wan, L., Zhu, L., Shen, J., Fu, H., Liu, W., Qin, J.: Triple-cooperative video shadow detection. In: IEEE Conference

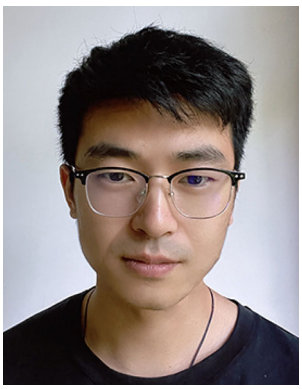
on Computer Vision and Pattern Recognition, pp. 2714–2723 (2021)
11. Chen, Z., Zhu, L., Wan, L., Wang, S., Feng, W., Heng, P.A.: A multi-task mean teacher for semi-supervised shadow detection. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 5610–5619 (2020)
12. Cheng, Z., Yang, Q., Sheng, B.: Deep colorization. In: IEEE International Conference on Computer Vision, pp. 415–423 (2015)
13. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Li, F.F.: ImageNet: a large-scale hierarchical image database. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255 (2009)
14. Deshpande, A., Lu, J., Yeh, M.C., Chong, M.J., Forsyth, D.: Learning diverse image colorization. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2877–2885 (2017)
15. Deshpande, A., Rock, J., Forsyth, D.: Learning large-scale automatic image colorization. In: IEEE International Conference on Computer Vision, pp. 567–575 (2015)
16. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: transformers for image recognition at scale. In: International Conference on Learning Representations, pp. 1–21 (2021)
17. Fang, F., Wang, T., Zeng, T., Zhang, G.: A superpixel-based variational model for image colorization. IEEE Trans. Vis. Comput. Graph. **26**(10), 2931–2943 (2020)
18. Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. In: International Conference on Artificial Intelligence and Statistics, vol. 9, pp. 249–256 (2010)
19. Gupta, R.K., Chia, A.Y.S., Rajan, D., Ng, E.S., Huang, Z.: Image colorization using similar images. In: ACM International Conference on Multimedia, pp. 369–378 (2012)
20. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: IEEE International Conference on Computer Vision, pp. 2980–2988 (2017)
21. He, M., Chen, D., Liao, J., Sander, P..V., Yuan, L.: Deep exemplar-based colorization. ACM Trans. Graph. **37**(4), 47:1-47:16 (2018)
22. Hou, L., Vicente, T.F.Y., Hoai, M., Samaras, D.: Large scale shadow annotation and detection using lazy annotation and stacked CNNs. IEEE Trans. Pattern Anal. Mach. Intell. **43**(4), 1337–1351 (2021)
23. Hou, Y., Zheng, L.: Multiview detection with shadow transformer (and view-coherent data augmentation). In: ACM International Conference on Multimedia, pp. 1673–1682 (2021)
24. Hu, X., Wang, T., Fu, C.W., Jiang, Y., Wang, Q., Heng, P.A.: Revisiting shadow detection: a new benchmark dataset for complex world. IEEE Trans. Image Process. **30**, 1925–1934 (2021)
25. Huang, Z., Zhao, N., Liao, J.: UniColor: a unified framework for multi-modal colorization with transformer. ACM Trans. Graph. **41**(6), 205:1-205:16 (2022)
26. Iizuka, S., Simo-Serra, E., Ishikawa, H.: Let there be color!: joint end-to-end learning of global and local image priors for automatic image colorization with simultaneous classification. ACM Trans. Graph. **35**(4), 110:1-110:11 (2016)
27. Irony, R., Cohen-Or, D., Lischinski, D.: Colorization by example. In: Eurographics Symposium on Rendering, pp. 201–210 (2005)
28. Ji, X., Jiang, B., Luo, D., Tao, G., Chu, W., Xie, Z., Wang, C., Tai, Y.: ColorFormer: image colorization via color memory assisted hybrid-attention transformer. In: European Conference on Computer Vision, pp. 20–36 (2022)
29. Kang, X., Yang, T., Ouyang, W., Ren, P., Li, L., Xie, X.: DDColor: towards photo-realistic image colorization via dual decoders. In: IEEE International Conference on Computer Vision, pp. 328–338 (2023)
30. Kim, E., Lee, S., Park, J., Choi, S., Seo, C., Choo, J.: Deep edge-aware interactive colorization against color-bleeding effects. In: IEEE International Conference on Computer Vision, pp. 14,647–14,656 (2021)
31. Kim, G., Kang, K., Kim, S., Lee, H., Kim, S., Kim, J., Baek, S.H., Cho, S.: BigColor: colorization using a generative color prior for natural images. In: European Conference on Computer Vision, pp. 350–366 (2022)
32. Kumar, M., Weissenborn, D., Kalchbrenner, N.: Colorization transformer. In: International Conference on Learning Representations, pp. 1–24 (2021)
33. Larsson, G., Maire, M., Shakhnarovich, G.: Learning representations for automatic colorization. In: European Conference on Computer Vision, pp. 577–593 (2016)
34. Levin, A., Lischinski, D., Weiss, Y.: Colorization using optimization. ACM Trans. Graph. **23**(3), 689–694 (2004)
35. Liu, L., Prost, J., Zhu, L., Papadakis, N., Liò, P., Schönlieb, C.B., Aviles-Rivero, A.I.: Scotch and soda: a transformer video shadow detection framework. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 10,449–10,458 (2023)
36. Liu, X., Wan, L., Qu, Y., Wong, T.T., Lin, S., Leung, C.S., Heng, P.A.: Intrinsic colorization. ACM Trans. Graph. **27**(5), 152:1-152:9 (2008)
37. Lu, X., Cao, Y., Liu, S., Long, C., Chen, Z., Zhou, X., Yang, Y., Xiao, C.: Video shadow detection via spatio-temporal interpolation consistency training. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 3106–3115 (2022)
38. Messaoud, S., Forsyth, D., Schwing, A.G.: Structural consistency and controllability for diverse colorization. In: European Conference on Computer Vision, pp. 603–619 (2018)
39. Qu, Y., Wong, T.T., Heng, P.A.: Manga colorization. ACM Trans. Graph. **25**(3), 1214–1220 (2006)
40. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representations, pp. 1–14 (2015)
41. Su, J.W., Chu, H.K., Huang, J.B.: Instance-aware image colorization. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 7965–7974 (2020)
42. Sýkora, D., Dingliana, J., Collins, S.: Lazybrush: flexible painting tool for hand-drawn cartoons. Comput. Graph. Forum **28**(2), 599–608 (2009)
43. Vasluianu, F.A., Seizinger, T., Timofte, R.: WSRD: a novel benchmark for high resolution image shadow removal. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 1826–1835 (2023)
44. Wang, J., Li, X., Yang, J.: Stacked conditional generative adversarial networks for jointly learning shadow detection and shadow removal. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1788–1797 (2018)
45. Wang, Y., Xia, M., Qi, L., Shao, J., Qiao, Y.: PalGAN: image colorization with palette generative adversarial networks. In: European Conference on Computer Vision, pp. 271–288 (2022)
46. Welsh, T., Ashikhmin, M., Mueller, K.: Transferring color to greyscale images. ACM Trans. Graph. **21**(3), 277–280 (2002)
47. Weng, S., Sun, J., Li, Y., Li, S., Shi, B.: CT^2: colorization transformer via color tokens. In: European Conference on Computer Vision, pp. 1–16 (2022)
48. Wu, Y., Wang, X., Li, Y., Zhang, H., Zhao, X., Shan, Y.: Towards vivid and diverse image colorization with generative color prior. In: IEEE International Conference on Computer Vision, pp. 14,357–14,366 (2021)
49. Xia, M., Hu, W., Wong, T.T., Wang, J.: Disentangled image colorization via global anchors. ACM Trans. Graph. **41**(6), 204:1-204:13 (2022)
50. Xiao, C., Han, C., Zhang, Z., Qin, J., Wong, T.T., Han, G., He, S.: Example-based colourization via dense encoding pyramids. Comput. Graph. Forum **39**(1), 20–33 (2020)
51. Xiao, Y., Wu, J., Zhang, J., Zhou, P., Zheng, Y., Leung, C.S., Kavan, L.: Interactive deep colorization and its application for

image compression. IEEE Trans. Vis. Comput. Graph. **28**(3), 1557–1572 (2022)
52. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: SegFormer: simple and efficient design for semantic segmentation with transformers. In: Advances in Neural Information Processing Systems, vol. 34, pp. 12,077–12,090 (2021)
53. Xie, X., Zhou, P., Li, H., Lin, Z., Yan, S.: Adan: adaptive nesterov momentum algorithm for faster optimizing deep models. In: Advances in Neural Information Processing Systems Workshop, pp. 1–8 (2022)
54. Zhang, L., Li, C., Wong, T.T., Ji, Y., Liu, C.: Two-stage sketch colorization. ACM Trans. Graph. **37**(6), 261:1-261:14 (2018)
55. Zhang, R., Isola, P., Efros, A.A.: Colorful image colorization. In: European Conference on Computer Vision, pp. 649–666 (2016)
56. Zhang, R., Zhu, J.Y., Isola, P., Geng, X., Lin, A.S., Yu, T., Efros, A.A.: Real-time user-guided image colorization with learned deep priors. ACM Trans. Graph. **36**(4), 119:1-119:11 (2017)
57. Zhao, J., Han, J., Shao, L., Snoek, C.G.M.: Pixelated semantic colorization. Int. J. Comput. Vis. **128**, 818–834 (2020)
58. Zheng, Q., Qiao, X., Cao, Y., Lau, R.W.H.: Distraction-aware shadow detection. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 5162–5171 (2019)
59. Zhu, J., Samuel, K.G.G., Masood, S.Z., Tappen, M.F.: Learning to recognize shadows in monochromatic natural images. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 223–230 (2010)
60. Zou, C., Mo, H., Gao, C., Du, R., Fu, H.: Language-based colorization of scene sketches. ACM Trans. Graph. **38**(6), 233:1-233:16 (2019)

Xin Duan received the B.Eng. degree in software engineering from the Harbin Institute of Technology, Shenzhen, 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. degree in computing with The Hong Kong Polytechnic University, Hong Kong. Her current research interests include low-level vision and computer graphics.

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 pursuing the Ph.D. degree in fashion and textiles with The Hong Kong Polytechnic University, Hong Kong. His current research interests include AI for computer graphics, deep learning, line drawing, and fashion colorization.

Renjie Zhang received the B.Eng. degree in software engineering from the Sun Yat-sen University, Guangzhou, China, in 2019. He is currently pursuing the Ph.D. degree in computing with The Hong Kong Polytechnic University, Hong Kong. His current research interests include human skeleton establishment, neural architecture search, pose estimation, deep learning, and 3D human reconstruction.



Xin Wang received the B.Eng. degree in computer science and technology from the Dalian University of Technology, Dalian, China, in 2017. He is currently pursuing the Ph.D. degree in computing with The Hong Kong Polytechnic University, Hong Kong. His current research interests include image synthesis, deep learning, diffusion models, and computer graphics.

Ping Li 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 200 top-tier scholarly research articles, 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.