Diffusart: Enhancing Line Art Colorization with Conditional Diffusion Models

Hernan Carrillo[1] †, Michaël Clément[1], Aurélie Bugeau[1,2], Edgar Simo-Serra[3]
[1]Univ. Bordeaux, CNRS, Bordeaux INP, LaBRI, UMR 5800
[2] Institut Universitaire de France (IUF)
[3]Waseda University
† hernan.carrillo-lindado@labri.fr

Abstract

Colorization of line art drawings is an important task in illustration and animation workflows. However, this highly laborious process is mainly done manually, limiting the creative productivity. This paper presents a novel interactive approach for line art colorization using conditional Diffusion Probabilistic Models (DPMs). In our proposed approach, the user provides initial color strokes for colorizing the line art. The strokes are then integrated into the conditional DPM-based colorization process by means of a coupled implicit and explicit conditioning strategy to generates diverse and high-quality colorized images. We evaluate our proposal and show it outperforms existing state-of-the-art approaches using the FID, LPIPS and SSIM metrics.

1. Introduction

Line art colorization is a popular technique used in various fields, such as art, animation, and graphic design. It involves adding color to grayscale line drawings to make them visually appealing and expressive. However, this process is laborious as it is typically carried out manually for traditional animation, mainly using software illustration tools such as Photoshop, ClipStudio, and Krita. Automated colorization have the potential to significantly enhance an artist's workflow.

In recent years, various methods have explored user-guided automatic line art colorization using deep learning. In particular, Generative Adversarial Network (GANs) architectures have been proposed [3, 20, 26–28]. These methods couple color hints as input with learned color priors from large-scale datasets to colorize the line art images. GAN architectures can achieve impressive and high-quality outputs. However, certain issues remain problematic, for example, ensuring color consistency with user color inputs and reaching color harmony between small image regions. In addition, GAN architectures can be challenging to train due to instabilities [1, 7]. To overcome these issues, Diffusion Probabilistic Models (DPMs) [9, 22] propose a framework capable of generating high-fidelity images by training a U-Net [17] like generator architecture and sampling from a Markov chain. These methods have been applied to various computer vision problems such as image synthesis [5], super-resolution [16, 19], and automatic image colorization [18] achieving better qualitative and quantitative results than previous state-of-the-art GANs architectures.

Figure 1. Diffusart enables the colorization of line arts created by an artist l (left) using color scribbles s (center). On the right is the result of our proposal $\hat{x}_0$.

In this paper, we introduce a new user-guided line art colorization model based on a conditional diffusion model that achieves better results than state-of-the-art methods. In addition, we explore the use of a coupled implicit and explicit conditioning strategy on the diffusion model for this application. An in-depth evaluation corroborates the efficiency of our approach.

2. Related work

Automatic methods for line art colorization have been primarily classical image processing optimization-based, as described in [15, 23]. These approaches typically involve using image features such as pattern and intensity continuity to propagate color hints over regions. However, their effectiveness is limited when applied to complex line drawings

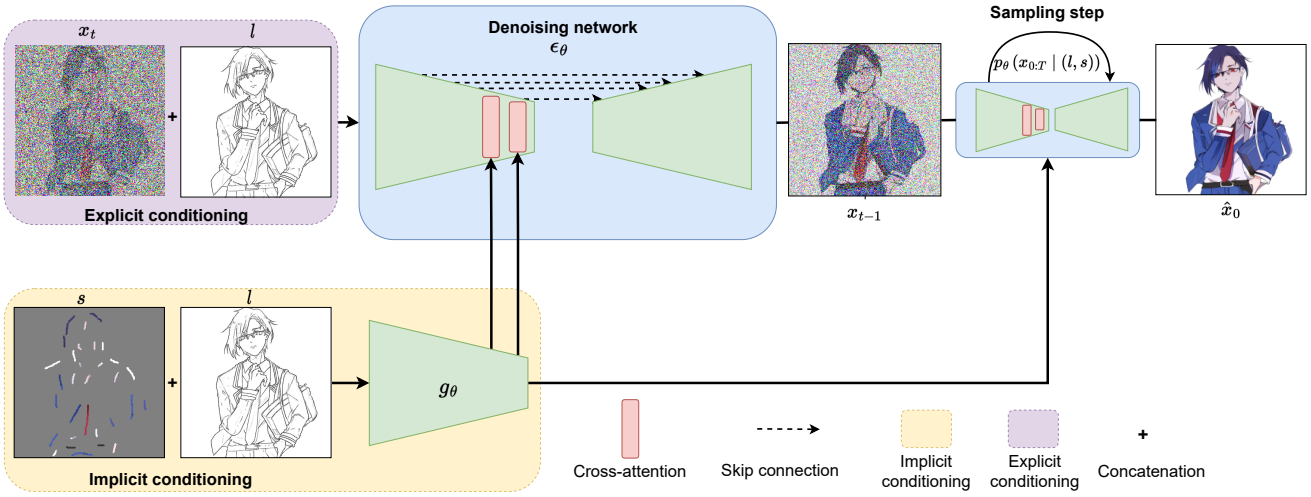

Figure 2. Overview of our proposed user-guided line art colorization. The framework is composed of two main components: a denoising model ϵ_θ, which learns to generate a denoised image, and an application-specific encoder g_θ for extracting user color scribbles information.

where a high amount of user color hints are needed.

Deep learning methods, and in particular GAN architectures [13, 27, 29], have been used for user-guided line art colorization where color scribbles are propagated based on neural networks trained with large amounts of data. In [3], Ci *et al*. propose a method to enhance the generalization capability of the neural network by introducing a local features network independent of synthetic data. Yliess *et al*. [26] improves the visual fidelity of results by using a double generator approach. Other methods explore the use of reference color images to transfer a particular artist style [6, 10, 12]. Although GANs can produce high-fidelity images, these architectures can be unstable to train which could produce perceptually unsatisfying results.

Diffusion Probabilistic Models (DPMs) [9, 22] seem to have the potential to overcome GANs issues in many applications. These methods have recently emerged as a class of generative models for high-dimensional data such as images and audio. DPMs transform the input data through a series of controlled noising/denoising steps to predict new data distributions. These methods have achieved state-of-the-art results in various image-to-image generation tasks, including image synthesis [5], super-resolution [11, 19], and colorization [18]. Inspired by these new methods, we propose a novel conditional diffusion model for line art colorization that can be guided by user color scribbles.

3. Proposed method

The objective is to colorize a grayscale line art image from user color scribbles. Our proposal uses a diffusion model, which learns to generate a colorized line art image $\hat{x}_0$ given grayscale line art l and color scribbles s (see Figure 1).

Our framework is composed of two main parts: a denoising model ϵ_θ from the main denoising pipeline and an application-specific encoder g_θ for extracting color scribbles information (see Figure 2). The first learns to denoise noisy images from an unknown distribution conditioned to a line art image l. The second part, g_θ, extracts color features previously inputted by the user to guide the line art colorization. Finally, the predicted image $\hat{x}_0$ is retrieved using the DDPM sampling algorithm [9].

Figure 3. Example of synthetic line art of a color image generate using SketchKeras [14], and Sketch Simplification [21].

3.1. Diffusion Models

Diffusion models, which have been used in various image generation applications, convert standard Gaussian distribution samples to empirical data distribution samples by employing a Markov chain denoising process of T steps. Given an initial image x_0 from the real distribution, the diffusion process successively adds Gaussian noise with vari-

ance β_t and mean $\mu_t = \sqrt{1-\beta_t}x_{t-1}$ to obtain intermediate noisy image x_t, this is called forward process. The idea is to learn a parametric approximation to the unknown conditional distribution $p_\theta(x_{t-1} \mid x_t)$ by utilizing a stochastic iterative refinement process. During inference, the aim is to reverse the Gaussian diffusion process through a reverse Markov chain according to a learned transition distribution p_θ:

$$p_\theta\left(x_{0:T}\right) = p_\theta\left(x_T\right) \prod_{t=1}^{T} p_\theta\left(x_{t-1} \mid x_t\right). \quad (1)$$

Finally, to learn the reverse chain, we use a neural denoising model ϵ_θ.

3.2. Conditioning Diffusion Models

Diffusion models such as those proposed by [9, 22] operate in a non-conditional setting (Equation (1)). Instead, we jointly use two approaches to condition our diffusion model. First, as in [16], we jointly train an application-specific encoder g_θ, which extracts semantic features from color scribbles and line art images. These features are then introduced to the denoising model ϵ_θ by means of cross-attention mechanism [24]. For the second approach, and inspired by [2, 18], we explicitly condition the predicted distribution on the denoising neural network ϵ_θ by directly concatenating line art image s to the noisy input x_t. Finally, by joining both conditioning in the inference process (1) changes to

$$p_\theta\left(x_{0:T} \mid (l,s)\right) = p_\theta\left(x_T\right) \prod_{t=1}^{T} p_\theta\left(x_{t-1} \mid x_t, (l,s)\right). \quad (2)$$

Given (2), the training process of our proposal on conditioned image pairs is trained using the L_1 loss as

$$L_1 = \mathbb{E}_{l,\epsilon\sim\mathcal{N}(0,1),t}\left[\left\|\epsilon - \epsilon_\theta\left(l, x_t, t, g_\theta(l,s)\right)\right\|_1\right], \quad (3)$$

where both the denoising model ϵ_θ and the encoder g_θ are jointly trained.

4. Experiments

4.1. Dataset preparation

Synthetic Line Art. We conducted experiments using a subsample of color illustrations from the dataset safe Danbooru2021 [4]. To filter out grayscale images, we utilized tags such as "grayscale" and "monochrome". We use 200k training images and 13k images for test. For creating synthetic line art, we rely on two extraction methods: SketchKeras [14] and Sketch simplification [21]. Figure 3 depicts an example of synthetic data generated by previously mentioned methods. Finally, the type of sketch at training time is randomly sample by a uniform distribution with a 50% probability of choosing SketchKeras or the Sketch simplification methods.

Simulated Color Scribbles. To achieve a model that can handle user color inputs, we simulate human scribbles by randomly sampling vertical, horizontal, and diagonal lines. We use three parameters: the number of scribbles sampled from the uniform distribution $\mathcal{U}(4, 25)$, scribble thickness sampled from $\mathcal{U}(1, 4)$ pixels, and scribble length sampled from $\mathcal{U}(5, 30)$ pixels. Additionally, as a high amount of illustrations in the dataset present white backgrounds, there is a high probability that the synthetic scribbles would bias the model toward the color white. Therefore, we only use the sampled synthetic scribbles when they contain less than 60% white pixels.

4.2. Implementation details

Our implementation was inspired by [9]. To reduce computational cost, we only use self-attention and cross-attention mechanisms on the bottleneck of the denoising model ϵ_θ. We use the Adam optimizer with a learning rate of $2e^{-5}$, a cosine warm-up schedule for 5k training steps, and a batch size of 40. We introduce a color feature extraction encoder g_θ that uses the same encoder architecture as ϵ_θ with only one Residual block per layer instead of two. Both the denoising model ϵ_θ and encoder g_θ are jointly trained from scratch. All line art and color scribble images are fixed to the resolution 256×256, and values are normalized to the range $[-1, 1]$. Our final method was trained for 80 epochs with 10 NVIDIA RTX 2080Ti GPUs.

Table 1. Quantitative comparison with state-of-the-art user-guided line art colorization methods.

Method	**SSIM↑**	**LPIPS ↓**	**FID↓**
AlacGAN [3]	0.66	0.26	13.27
PaintsTorch [26]	0.79	**0.14**	8.79
Ours (w/o explicit cond.)	0.77	0.15	7.91
Ours (full)	**0.81**	**0.14**	**6.15**

5. Experimental validation

To evaluate the effectiveness of our line art colorization framework, we compare our results quantitatively and qualitatively with two other state-of-the-art user-guided line art colorization approaches [3, 26]. In order to do a fair comparison, we retrained all the methods with the same dataset and used the default parameters presented in their methods.

Quantitative Evaluation. We use three metrics to compare different methods quantitatively. The first metric is the Structural Similarity (SSIM) [25], which examines the model's ability to reconstruct the content of the original image. The second metric is the Learned Perceptual Image Patch Similarity (LPIPS) [30], which measures perceptual similarities between two images and correlates strongly with

Figure 4. Comparison of our proposed method with different user-guided line art colorization methods: AlacGAN [3] and PaintsTorch [26].

human perception. The last metric, the Fréchet Inception Distance (FID) [8], is used to measure a perceptual similarity between two sets of images. All three metrics are calculated on 13k test images, between the color illustration image as ground-truth and generated images with fixed color hints from the different methods.

As shown in Table 1 our results retain 15% and 2% more structural information than the other two state-of-the-art methods. For the LPIPS metric, our method surpasses [3] and achieves comparable results to [26]. Finally, on the FID metric, we outperform both methods. In addition, using only implicit conditioning reduces the performance compared to our full method.

Qualitative Evaluation. Figure 4 shows the results from the two state-of-the-art methods [3], [26], and ours. We can see that our qualitative results are consistent with quantitative scores, showing more high-quality details, a visually appealing colorization, and a more accurate representation of color shades compared to the other two methods.

6. Conclusion

In this paper, we introduced a novel approach for user-guided line art colorization using conditional Diffusion Models. Our proposal exploits a coupled implicit and explicit conditioning strategy that ensures a robust structural generation of details and an accurate representation of user colors. Experimental evaluation on a large-scale dataset shows that our method outperforms existing techniques both quantitatively and qualitatively.

Acknowledgements. This study has been carried out with financial support from the French Research Agency through the PostProdLEAP project (ANR-19-CE23-0027-01).

References

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, 2017. 1

[2] Nanxin Chen, Yu Zhang, Heiga Zen, Ron J Weiss, Mohammad Norouzi, and William Chan. Wavegrad: Estimating gradients for waveform generation. In *International Conference on Learning Representations*, 2021. 3

[3] Yuanzheng Ci, Xinzhu Ma, Zhihui Wang, Haojie Li, and Zhongxuan Luo. User-guided deep anime line art colorization with conditional adversarial networks. In *ACM International Conference on Multimedia*, 2018. 1, 2, 3, 4

[4] DanbooruCommunity. Danbooru2021: A large-scale crowdsourced and tagged anime illustration dataset., 2021. 3

[5] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 2021. 1, 2

[6] Chie Furusawa, Kazuyuki Hiroshiba, Keisuke Ogaki, and Yuri Odagiri. Comicolorization: semi-automatic manga colorization. In *International Conference on Computer Graphics and Interactive Techniques*, pages 1–4, 2017. 2

[7] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved training of wasserstein gans. In *Advances in Neural Information Processing Systems*, 2017. 1

[8] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, 2017. 4

[9] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2020. 1, 2, 3

[10] Junsoo Lee, Eungyeup Kim, Yunsung Lee, Dongjun Kim, Jaehyuk Chang, and Jaegul Choo. Reference-based sketch image colorization using augmented-self reference and dense semantic correspondence. In *Conference on Computer Vision and Pattern Recognition*, June 2020. 2

[11] Haoying Li, Yifan Yang, Meng Chang, Shiqi Chen, Huajun Feng, Zhihai Xu, Qi Li, and Yueting Chen. Srdiff: Single image super-resolution with diffusion probabilistic models. *Neurocomputing*, pages 47–59, 2022. 2

[12] Zekun Li, Zhengyang Geng, Zhao Kang, Wenyu Chen, and Yibo Yang. Eliminating gradient conflict in reference-based line-art colorization. In *European Conference on Computer Vision*, pages 579–596, 2022. 2

[13] Yifan Liu, Zengchang Qin, Tao Wan, and Zhenbo Luo. Auto-painter: Cartoon image generation from sketch by using conditional wasserstein generative adversarial networks. *Neurocomputing*, 311:78–87, 2018. 2

[14] lllyasviel. sketchkeras, 2017. 2, 3

[15] Yingge Qu, Tien-Tsin Wong, and Pheng-Ann Heng. Manga colorization. *ACM Transactions on Graphics*, 25(3), 2006. 1

[16] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Conference on Computer Vision and Pattern Recognition*, 2022. 1, 3

[17] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention*, pages 234–241, 2015. 1

[18] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM International Conference on Multimedia*, 2022. 1, 2, 3

[19] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 1, 2

[20] Patsorn Sangkloy, Jingwan Lu, Chen Fang, Fisher Yu, and James Hays. Scribbler: Controlling deep image synthesis with sketch and color. In *Conference on Computer Vision and Pattern Recognition*, 2017. 1

[21] Edgar Simo-Serra, Satoshi Iizuka, and Hiroshi Ishikawa. Mastering Sketching: Adversarial Augmentation for Structured Prediction. *Transactions on Graphics*, 2018. 2, 3

[22] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems*, 2019. 1, 2, 3

[23] Daniel Sýkora, John Dingliana, and Steven Collins. Lazybrush: Flexible painting tool for hand-drawn cartoons. *Computer Graphics Forum*, 28, 2009. 1

[24] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3

[25] Zhou 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*, 2004. 3

[26] Hati Yliess, Gergor Jouet, Francis Rousseaux, and Clement Duhrat. Paintstorch: A user-guided anime line art colorization tool with double generator conditional adversarial network. In *ACM Eur. Conf. Visual Media Production*, 2019. 1, 2, 3, 4

[27] Mingcheng Yuan and Edgar Simo-Serra. Line Art Colorization with Concatenated Spatial Attention. In *Conference on Computer Vision and Pattern Recognition Workshops*, 2021. 1, 2

[28] Lvmin Zhang, Chengze Li, Edgar Simo-Serra, Yi Ji, Tien-Tsin Wong, and Chunping Liu. User-Guided Line Art Flat Filling with Split Filling Mechanism. In *Conference on Computer Vision and Pattern Recognition*, 2021. 1

[29] Lvmin Zhang, Chengze Li, Tien-Tsin Wong, Yi Ji, and Chunping Liu. Two-stage sketch colorization. *ACM Transactions on Graphics*, 2018. 2

[30] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Conference on Computer Vision and Pattern Recognition*, 2018. 3