Enhancing Anime Line Drawing Colorization: With AnimeDiffusion Inspired Diffusion Models

Tanbir Yousuf 21-44394-1
Yulad Hassan 20-43776-2
Pritom Chandra Dey 21-44407-1
Afrin Saidatun Neshat 21-44384-1

Abstract: This report discusses AnimeDiffusion, a novel model designed for the automated colorization of anime face line drawings. The model employs a diffusion approach, integrating a U-Net architecture with attention blocks and the Extended Difference of Gaussians (XDoG) algorithm for line extraction. A two-stage training strategy, involving classifier-free guidance pre-training and image reconstruction guidance fine-tuning, enhances the model's proficiency. The researchers present a benchmark dataset and conduct comprehensive experiments, showcasing AnimeDiffusion's superior performance over existing Generative Adversarial Networks (GANs)-based models. The paper identifies limitations related to color gaps and stylistic disparities and proposes future directions, including multi-modal input colorization and improved training techniques.

Keywords: AnimeDiffusion, line drawing colorization, diffusion models, U-Net architecture, XDoG algorithm, benchmark dataset, two-stage training, generative adversarial networks, multi-modal input, computer vision

I. Introduction

In the realm of animation, the colorization of line drawings stands as a vital but time-consuming process, particularly when dealing with intricate content structures. Line drawing colorization is a technique that involves adding color to black and white line drawing. Manually colorizing the line drawings is time consuming, especially for the line drawings with complex structure content. The development of an automated line drawing colorization system is more efficient. Early attempts utilized neural networks for cartoon colorization, refining results often required numerous interactions. [2] Subsequent user-hint-based methods aimed to enhance control but remained inconvenient, especially for amateur users. Reference-based colorization methods [3] provided a more user-friendly alternative, requiring only a line drawing and a corresponding reference color image. However, existing approaches relying on generative adversarial networks (GANs) faced challenges, such as instability during training due to the deployment of multiple losses. These approaches mainly focus on the improvement of feature aggregation module of two extracted deep features. To overcome these limitations, this paper introduces AnimeDiffusion, a diffusion model tailored for anime face line drawing colorization. Inspired by the efficiency of diffusion probabilistic models [4], this paper proposes to design a hybrid training strategy involving classifier-free guidance pre-training and image reconstruction guidance fine-tuning. Notably, AnimeDiffusion's fine-tuning allows for more efficient training from sketch to colorization, showcasing superior results compared to state-of-the-art GANs-based models. To facilitate further research, a novel benchmark dataset is introduced, derived from high-resolution anime face images. This comprehensive approach positions AnimeDiffusion as a noteworthy advancement in the automation of anime line drawing colorization within the animation industry.

Figure 1: AnimeDiffusion: Anime Face Line Drawing Colorization via Diffusion Models [1].

The main objective of this paper can be summarized as follows:

- Aim to use the benchmark dataset for creating scarcity of high-resolution anime face datasets necessary for evaluating line drawing colorization algorithms.
- Comprehensive experiments and a user study affirming the superior performance of AnimeDiffusion, both qualitatively and quantitatively, in comparison to cutting-edge GANs-based methods.
- The distinctive ability of AnimeDiffusion to precisely colorize anime face line drawings featuring heterochromatic pupils without the need for specialized modules for eyes or pupils.
- To explore multi-modal input line drawing colorization by integrating text information and reference images.

II. Related works:

A. Line Drawing Colorization

Traditional optimization-based approaches ([5], [6]) allow users to use brushes to inject desired color into specific regions of line drawings. With the advancement of deep learning technology, user-hint colorization methods have emerged. In these methods, color hints are concatenated with line drawings and serve as input for neural networks. Ci et al. proposed a conditional GAN model for anime line drawing

colorization using color scribbles ([7]). Zhang et al. developed a two-stage colorization method with color points hints, dividing the complex task into simpler subtasks ([8]). Kim et al. utilized their SECat module with text tags for detailed illustrations ([9]). Zou et al. introduced a language-based system for interactive colorization of scene sketches ([10]). However, user-hint methods pose challenges as they become more labor-intensive with an increasing number of line drawings, requiring multiple interactions for result refinement. These methods are not user-friendly for amateur users without aesthetic judgment, especially in preparing appropriate color hints. To address these challenges, various reference-based colorization methods have been proposed. AnimeDiffusion method is introduced as a novel reference-based colorization tailored for anime face line drawing colorization. It claims to generate better results in both visual quality and quantitative metrics compared to previous GAN-based methods, while maintaining fidelity to the reference image.

B. Semantic Correspondence

Semantic correspondence is a critical challenge in computer vision, aiming to establish dense correspondence across images with similar semantic content. In computer graphics, it is crucial for exemplar-based image colorization tasks. Zhang et al. [11] proposed an exemplar-based image translation system, and Lee et al. [12] introduced an attention-based module for spatially matching and aggregating sketch features with reference color image features. Other methods include Li et al. [13], He et al. [14], Zhang et al. [15], and Lu et al. [16], which focus on semantic correspondence learning for grayscale image or video colorization. However, the AnimeDiffusion method is designed specifically for anime face line drawing colorization, generating results with clear and accurate semantic colors.

C. Diffusion Models

Diffusion models, such as Denoising Diffusion Probabilistic Models (DDPM) and Denoising Diffusion Implicit Models (DDIM), have shown success in image generation tasks. These models offer stable training and high-quality generation. Zhang et al. [17] proposed ControlNet, which can generate diverse colored cartoon images based on sketch input and text prompt, utilizing diffusion models pre-trained on natural image datasets. However, AnimeDiffusion is highlighted as the first method to perform reference-based anime face line drawing colorization using diffusion models, achieving accurate color information without relying on prior knowledge from natural image datasets. Here is an example of DDPM done by the team for a understanding of the model.

III. Model and Implementation:

A. Model Architecture:

AnimeDiffusion uses a U-Net architecture with attention blocks. The U-Net architecture is a type of convolutional neural network that is widely used for image segmentation tasks. The attention blocks within the network help it focus on important features in the input image. The model is trained to predict the noise added to the original color image, which is a unique approach compared to traditional colorization methods. Figure 2 shows the pipeline of the model.

B. Line Extraction:

The line extraction module uses the XDoG (Extended Difference of Gaussians) algorithm. This algorithm is a more advanced version of the traditional edge detection algorithm called Difference of Gaussians (DoG). The XDoG algorithm is optimized to extract lines that closely resemble those drawn by a professional artist, which enhances the quality of the line drawings used for colorization.

C. Training Strategy:

The training strategy of AnimeDiffusion consists of two stages. The first stage is a pre-training stage where the model learns to predict the noise added to the original color image. The second stage is a fine-tuning stage where the model learns to reconstruct the original color image from the noisy image. This two-stage training strategy helps the model learn a robust representation of the colorization task.

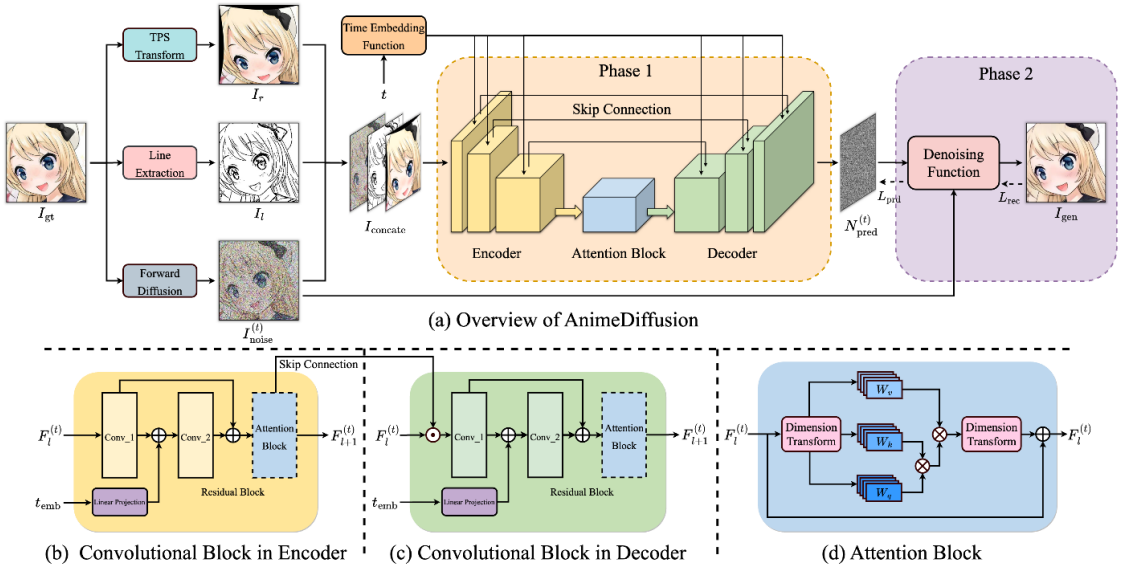


Figure 2: The flowchart of AnimeDiffusion [1]

D. Dataset:

The authors collected a new anime face line drawing colorization benchmark dataset, which contains 31696 training data and 579 testing data. The dataset is composed of high-resolution (256 × 256) anime face images with various styles and colors. The authors also applied XDoG extractor and TPS transformation to obtain line drawings and distorted reference images.

E. Implementation Details:

The authors implemented their AnimeDiffusion model based on the PyTorch framework and trained it on a NVIDIA A100 GPU. They used a linear noise schedule of (1e−6, 1e −2) with 1000 time steps, and applied the Adam optimizer with a learning rate of 1e −5. They pre-trained the model with a batch size of 32 for 300 epochs and fine-tuned the model with a batch size of 4 for 1 epoch.

F. Evaluation Methods:

The authors used three evaluation metrics to compare their model with other state-of-the-art GANs-based methods: FID, PSNR, and MS-SSIM. They also conducted a user study to assess the visual quality and preference of the generated images. They designed two kinds of colorization tasks: self-reference reconstruction and random-reference colorization.

G. Studies performed:

The authors performed qualitative and quantitative evaluations to demonstrate the effectiveness and superiority of their model. They showed that their model can generate high-quality images with accurate color and semantic correspondence and handle challenging cases such as heterochromatic pupils. They also performed ablation experiments to verify the impact of their hybrid training strategy. They showed that their fine-tuning stage can improve the colorization performance and reduce the computation cost. They also collaborated with professional artists to test and apply their model for original anime character colorization.

Figure 3 shows the 2 different eye color cases compared to 4 other projects.

Figure 3: Qualitative comparison for anime face with homochromatic and heterochromatic pupils. (a) reference images, (b) line drawings, (c) Lee et al. [11], (d) Li et al. [12], (e) Cao et al. [18], (f) Xu et al., and (g) AnimeDiffusion [1].

IV. Key Findings:

In this section, we will talk about the drawbacks of this model, which results in some inconsistencies for accurate results.

A. Problem 1:

Fig. 8. Reference-based line drawing colorization test for ablation study. (a) Reference images, (b) Results without fine-tuning, (c) Results with fine-tuning for 1 epoch, (d) Results with fine-tuning for 10 epochs.

For the reference-based line drawing colorization test, we show results in Fig. 8. Although the model without finetuning can distinguish regions that need different colors, there is still a color gap between the generated images and the reference images, and the result-colored image looks dimmer. According to the authors, training the diffusion model with classifier-free guidance is time-consuming, so they briefly fine-tuned the model and got much better results.

This nuanced adjustment not only underscored the efficacy of targeted model refinement but also highlighted the judicious balance between computational efficiency and quality augmentation. The fine-tuning process stands as a testament to the pragmatic amalgamation of theoretical prowess and practical considerations, ultimately culminating in superior output quality within a resource-constrained framework. Such strategic interventions illuminate the iterative nature of model development, emphasizing the pivotal role of fine-tuning in bridging the gap between model-generated outputs and desired reference benchmarks.

B. Problem 2:

Fig. 9 Limitation of our approach. (a) Reference images, (b) Colorization results, (c) Ground truth.

Due to the large stylistic differences between the reference image and the line drawing, the color of the mouth is not available in the reference image Fig. 9, while the teeth are not correctly identified in our model in the line drawing, and the color of the teeth is not accurately reflected in our colorization results.

There is a limitation in our method. Our model uses paired training data in training, and there is some style correlation between the reference image and the line drawings. For special style line drawings, such as Chibi cartoons as shown in Fig. 9(c), if the corresponding semantic information does not exist in the reference image, the colorization result of the line drawing may appear to be inconsistent with the real image.

V. What the researchers suggest

There is no doubt that the anime diffusion model will help animators reduce manual tasks and increase efficiency. Previously we have discovered a few drawbacks also. However, we know there is always a place for improvement.

In the future researchers are willing to work on multi-modal input line drawing colorization such as combining text information and reference images together to make the interactive way of colorization richer. This will greatly reduce the manual tasks of animators and improve the creation efficiency and colorization effect of the animation creation industry.

Multi-Modal Input Line Drawing Colorization

This could involve incorporating text information or other forms of guidance in addition to reference images to improve the colorization process. By combining different modalities, it may be possible to enhance the consistency and accuracy of colorization results, even in cases where the reference image lacks certain semantic information.

The advantage of employing multi-modal inputs lies in their ability to compensate for the inherent ambiguity present in line drawings. By integrating multiple data sources, the model can make more informed decisions about color placement, resulting in more accurate and vivid colorizations.

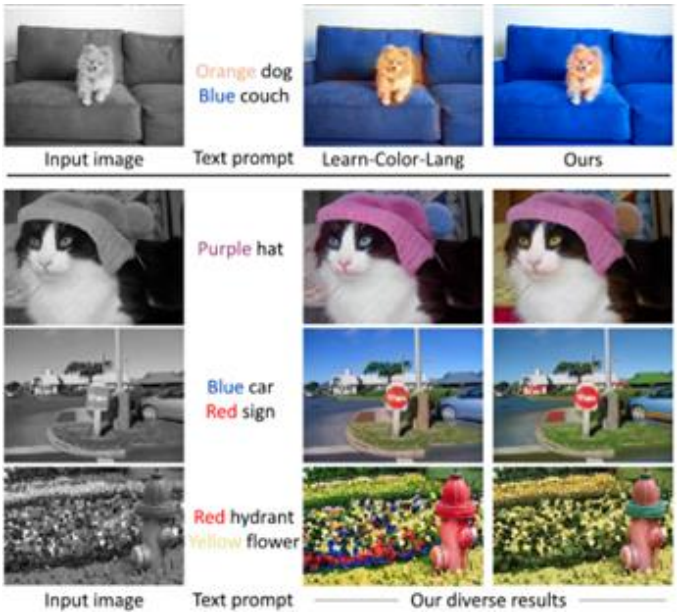
This approach showcases the versatility of AI models by enabling them to learn from various data modalities, thereby enhancing their ability to perform complex tasks like image colorization with greater accuracy and detail.

VI. What we suggest

In this section, we will suggest some techniques that might help the model to be more consistent and increase accuracy based on what we have learned throughout our entire CVPR course.

A. More Training

According to the authors the model needs more training to generate results with accurate color information. Proper training is crucial in various fields and contexts, particularly in machine learning.



B. Augmenting Datasets with Diverse Style

Expand the training dataset to include a more diverse range of artistic styles, especially those significantly different from the existing correlations between reference images and line drawings.

C. Unpaired or Weakly Supervised Learning

This could be the game changer for this model. If we manage to train the model using unpaired image-to-image translation techniques or self-supervised learning methods this will decrease dependency on paired learning so that the model can learn more precisely. Using unsupervised or weakly supervised learning approaches that don't rely solely on paired data could be a better option.

D. Style Disentanglement Techniques:

Implement techniques that disentangle style information from content information in the dataset, allowing the model to focus on learning style-independent representations.

Implementing one or a combination of these solutions can help mitigate the issue of style correlation between reference images and line drawings, enabling the Anime Diffusion model to generate more consistent and accurate images across a wider range of styles present in anime artwork.

VII. Conclusion

In conclusion, the AnimeDiffusion model marks a notable advancement in the automation of anime line drawing colorization. The introduced diffusion model, employing a U-Net architecture with attention blocks and XDoG line extraction, showcases effectiveness in predicting noise and reconstructing original color images through a two-stage training strategy.

Acknowledging limitations, the researchers emphasize the significance of fine-tuning to address color gaps and stylistic disparities between reference images and line drawings. Future directions include exploring multi-modal input colorization and implementing improvements such as more extensive training, diverse dataset augmentation, and unpaired or weakly supervised learning.

Further refinement is suggested through the exploration of style disentanglement techniques, aiming to enhance the model's consistency and accuracy. Continuous improvement remains crucial in the dynamic field of computer vision, underscoring the iterative nature of model development and adaptation to emerging techniques.

Reference

[1] Y. Cao, X. Meng, P. Y. Mok, X. Liu, T.-Y. Lee, and P. Li, 'AnimeDiffusion: Anime Face Line Drawing Colorization via Diffusion Models', arXiv [cs.CV]. 2023.

[2] D. Varga, C. A. Szabo, and T. Sziranyi, "Automatic cartoon colorization based on convolutional neural network," in Proceedings of the 15th International Workshop on Content-Based Multimedia Indexing, 2017, pp. 1–6.

[3] 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 Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 5801–5810.

[4] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in International Conference on Machine Learning. PMLR, 2015, pp. 2256–2265.

[5] Y. Qu, T.-T. Wong, and P.-A. Heng, “Manga colorization,” ACM Transactions on Graphics (TOG), vol. 25, no. 3, pp. 1214–1220, 2006.

[6] D. Sykora, J. Dingliana, and S. Collins, “Lazybrush: Flexible paint- ` ing tool for hand-drawn cartoons,” in Computer Graphics Forum, vol. 28, no. 2. Wiley Online Library, 2009, pp. 599–608.

[7] Y. Ci, X. Ma, Z. Wang, H. Li, and Z. Luo, “User-guided deep anime line art colorization with conditional adversarial networks,” in Proceedings of the 26th ACM international conference on Multimedia, 2018, pp. 1536–1544.

[8] L. Zhang, C. Li, T.-T. Wong, Y. Ji, and C. Liu, “Two-stage sketch colorization,” ACM Transactions on Graphics (TOG), vol. 37, no. 6, pp. 1–14, 2018.

[9] H. Kim, H. Y. Jhoo, E. Park, and S. Yoo, “Tag2pix: Line art colorization using text tag with secat and changing loss,” in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 9056–9065.

[10] C. Zou, H. Mo, C. Gao, R. Du, and H. Fu, “Language-based colorization of scene sketches,” ACM Transactions on Graphics (TOG), vol. 38, no. 6, pp. 1–16, 2019.

[11] P. Zhang, B. Zhang, D. Chen, L. Yuan, and F. Wen, “Cross-domain correspondence learning for exemplar-based image translation,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 5143–5153.

[12] J. Lee, E. Kim, Y. Lee, D. Kim, J. Chang, and J. Choo, “Referencebased sketch image colorization using augmented-self reference and dense semantic correspondence,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 5801–5810.

[13] H. Li, B. Sheng, P. Li, R. Ali, and C. P. Chen, “Globally and locally semantic colorization via exemplar-based broad-gan,” IEEE Transactions on Image Processing, vol. 30, pp. 8526–8539, 2021.

[14] M. He, J. Liao, D. Chen, L. Yuan, and P. V. Sander, “Progressive color transfer with dense semantic correspondences,” ACM Transactions on Graphics (TOG), vol. 38, no. 2, pp. 1–18, 2019.

[15] B. Zhang, M. He, J. Liao, P. V. Sander, L. Yuan, A. Bermak, and D. Chen, “Deep exemplar-based video colorization,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 8052–8061.

[16] P. Lu, J. Yu, X. Peng, Z. Zhao, and X. Wang, “Gray2colornet: Transfer more colors from reference image,” in Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 3210– 3218.

[17] L. Zhang and M. Agrawala, “Adding conditional control to textto-image diffusion models,” 2023.

[18] Y. Cao, H. Tian, and P. Mok, “Attention-aware anime line drawing colorization,” arXiv preprint arXiv:2212.10988, 2022.