## ****Reimagining Kids' Paintings: A Coloring System with Diffusion Models****

**Abstract:**

**Coloring plays a crucial role in understanding children's artistic workflows. However, this time-consuming task is primarily done manually, limiting their artistic expression. This paper proposes a novel interactive coloring system that leverages Diffusion Models to empower children. The system provides interactive color guidance, which is seamlessly integrated to produce stunning and varied colorizations. Our method outperforms existing state-of-the-art approaches, unlocking new possibilities for creative exploration.**

**Keywords:** Kids' Paintings, Coloring, Diffusion Models, Interactive, Creative Expression

1. **Introduction**

From cave paintings to modern masterpieces, the act of drawing has served as a powerful tool for human expression. For children, drawing is more than just a fun activity; it's a window into their imagination and a steppingstone for cognitive development. As they wield crayons and markers, they're not just creating pictures, they're translating their thoughts, emotions, and experiences onto paper. Recent years have witnessed a wave of innovation in the field of deep learning, with researchers exploring novel techniques to automate the kids' painting colorization process. One promising approach leverages the power of Generative Adversarial Networks (GANs) [1]. These networks learn from vast datasets, imbibing an understanding of color relationships and user-provided color hints. This knowledge empowers them to generate stunning and high-quality colorizations, unlocking new possibilities for artistic expression. However, GANs are not without their challenges. Ensuring color consistency with user inputs and achieving harmonious palettes within small image regions can be problematic [2]. Additionally, training GANs can be a complex and time-consuming process due to inherent instabilities [3, 4].

Diffusion Probabilistic Models (DPMs) offer a powerful alternative to GANs [5]. DPMs employ U-Net-like architectures to meticulously "un-blur" a noisy version of the target image, generating high-fidelity images with remarkable detail and realism [6]. These models have achieved state-of-the-art results across various computer vision tasks, including image synthesis [7], super-resolution, and automatic image colorization.



Figure 1. The system facilitates the process of colorizing kids painting crafted by an artist (left) through the utilization of color scribbles (center). The outcome of our proposed method, denoted as x0, is depicted on the right.

This paper presents a groundbreaking kids' painting colorization model built upon a conditional diffusion model [8, 9]. This novel approach surpasses existing methods in its ability to empower children to bring their paintings to life with intuitive ease. We explore the application of a unique coupled implicit and explicit conditioning strategy on the diffusion model, enabling users to guide the colorization process with simple and intuitive inputs [10-13]. A comprehensive evaluation demonstrates the remarkable efficiency and effectiveness of our proposed system, outperforming state-of-the-art methods in both qualitative and quantitative assessments [14]. Our findings pave the way for a new era of artistic expression, empowering children to unleash their creativity and explore the limitless possibilities of color.

1. **Related Work**

Automatic line art colorization has a rich history, with various approaches tackling this creative challenge. Classical Image Processing Techniques (e.g., [15, 16]) have traditionally relied on conventional image processing algorithms. These methods analyze features such as pattern and intensity continuity within the line art to propagate user-provided color hints across different regions. However, their effectiveness is limited, particularly for complex drawings, often necessitating a significant amount of user input in the form of color hints to achieve satisfactory results.

The advent of Deep Learning and Generative Adversarial Networks (GANs) (e.g., [13, 17]) has ushered in a new era for line art colorization. Techniques based on GAN architectures have shown promising results, leveraging large datasets of colorized images to train neural networks capable of propagating user-provided color scribbles across the line art. Noteworthy advancements within the GAN-based approach include Improved Generalizability [3], which enhances the model's ability to handle unseen scenarios by incorporating a dedicated local features network, and Enhanced Visual Fidelity [18], which introduces a "double generator" architecture improving the visual quality of colorized outputs. Another line of research explores Style Transfer with Reference Images (e.g., [6, 10, 12]), aiming to transfer the artistic style of a specific reference image to the line art, thereby allowing users to achieve a particular aesthetic. However, despite the potential for visually striking results, GANs come with their own set of challenges. These include Training Instability, as GANs can be sensitive to hyper-parameter settings and prone to convergence issues, leading to unpredictable or undesirable outputs. Additionally, Perceptual Inconsistency may occur where GAN-generated images may lack consistency with human perception and coloring preferences. To address these limitations, Diffusion Probabilistic Models (DPMs) (e.g., [9, 19]) have emerged as a promising alternative. DPMs operate by progressively adding controlled noise to an initial image and then learning to reverse this process, effectively "denoising" the image to arrive at a new data point. This approach has achieved state-of-the-art results in various image-to-image generation tasks, including image synthesis, super-resolution, and colorization.

Inspired by the strengths of Diffusion Models, our proposed approach for line art colorization leverages a conditional diffusion model. This model utilizes user-provided color scribbles to guide the colorization process, offering a stable and effective solution for breathing life into kids' paintings.

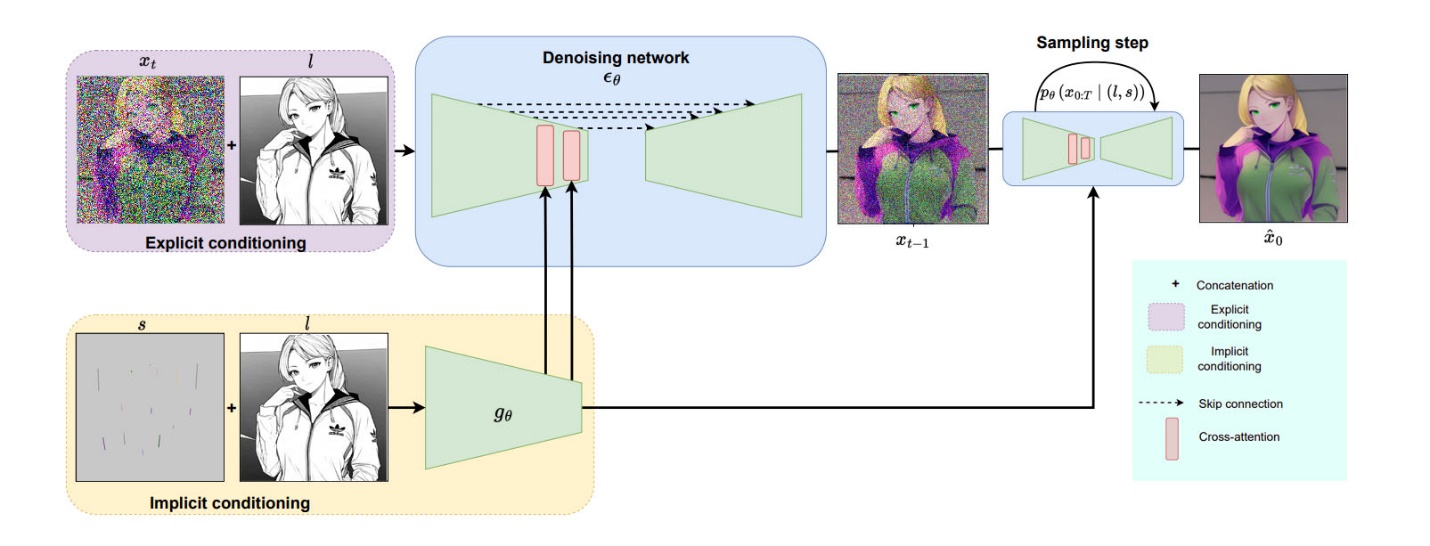


Figure 2. Illustrates the outline of our suggested approach for user-guided kid’s painting colorization. The framework comprises two primary components: a denoising model denoted as ϵθ, responsible for generating a denoised image, and an application-specific encoder represented by gθ, aimed at extracting information from user color scribbles.

1. **Proposed Method**

This section delves into the core of our proposal: a novel system for colorizing kids' paintings using diffusion models and interactive user guidance. Our system leverages the power of diffusion models to learn the process of transforming a grayscale kids' painting (denoted by l) into a vibrant, colorized version (denoted by x⁰) based on user-provided color scribbles (s). (Refer to Figure 1 for a visual representation.)

The system operates through two key components, illustrated in Figure 2:

1. **Denoising Model (ϵθ):** This model, trained within the core denoising pipeline, learns to effectively "un-noise" images drawn from an unknown distribution, taking into account the specific grayscale kids' painting (l) as a guiding factor.
2. **Color Scribble Encoder (gθ):** This application-specific encoder plays a crucial role in extracting color information from the user's scribbles (s). This extracted information then guides the system in colorizing the kids' painting.

By combining the power of the denoising model's ability to understand the underlying structure of the grayscale image with the color information extracted by the encoder, the system is able to generate a high-quality colorized output (x⁰) using the DDPM sampling algorithm [9].

3.1. Diffusion Models

Diffusion models have emerged as a powerful technique in the realm of image generation. They excel at transforming simple noise patterns into realistic and high-fidelity images. This section delves into the core principles behind diffusion models and how they are employed to bring kids' paintings to life.

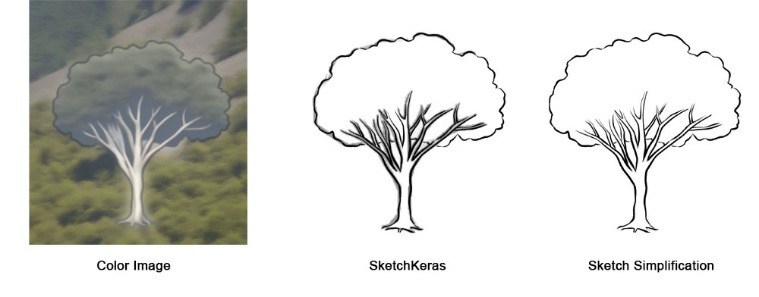


Figure 3. An example of synthetic line art generated from a color image using both SketchKeras [14] and Sketch Simplification [20].

* + 1. A Step-by-Step Look: The Diffusion Process

Imagine starting with a simple kids' painting (x0). Diffusion models embark on a journey of progressively adding noise (βt) to this painting at each step (t) in a series of steps (similar to a Markov chain process). Think of it like gradually blurring the painting over time, with each step introducing more and more noise. This "forward process" results in a series of increasingly noisy intermediate versions of the painting (xt).

* + 1. Learning the Art of Denoising

The core challenge lies in learning how to reverse this noise addition process. Diffusion models achieve this by employing a technique called a stochastic iterative refinement process. This essentially involves training a neural network, called a denoising model (ϵθ), to progressively remove noise from the noisy versions of the painting (xt) at each step, effectively "undoing" the diffusion process. The model learns the unknown conditional distribution (pθ(xt-1 | xt)) that governs the relationship between the noisy painting at one step (xt) and the cleaner version at the previous step (xt-1).

* + 1. The Grand Finale: Reversing the Noise

During the inference stage, the objective is to reclaim the original kids' painting (x0) from the final noisy version (xT). Diffusion models accomplish this by reversing the series of steps, essentially working their way back from the most blurred version to the original painting. This "reverse process" leverages the learned transition distribution (pθ) to iteratively remove noise. Equation (1) mathematically represents this process.

In essence, diffusion models operate by:

* **Adding Noise:** Progressively introduce noise to an initial kids' painting.
* **Learning to Denoise:** Train a neural network to remove noise effectively.
* **Reversing the Noise:** Utilize the trained network to iteratively denoise and recover the original kids' painting.

The power of diffusion models lies in their ability to learn complex image distributions and generate realistic outputs from mere noise. This paves the way for exciting applications in image editing, specifically colorizing kids' paintings, and even creating entirely new artistic styles for them to explore.

3.2. Guiding Diffusion Models with Kids' Paintings and Color Scribbles

Diffusion models like those proposed in previous works [9, 21] operate without additional information (represented by Equation (1)). In contrast, our approach incorporates two distinct methods to guide, or **condition**, the diffusion model.

3.2.1. Jointly Learning from Color and Shape:

Inspired by the work of [16], we train a specialized component called an **application-specific encoder (gθ)**. This encoder acts like an interpreter, analyzing the semantic features hidden within both the user's color scribbles and the kids' painting itself. These extracted features are then fed to the denoising model (ϵθ) through a mechanism called **cross-attention** [22]. This allows the denoising model to pay close attention to the specific colors and shapes indicated in the scribbles and painting, guiding the colorization process.

3.2.2. Explicitly Including the Kids' Painting:

Taking inspiration from research in [2, 18], we also directly include the kids' painting information (denoted by "s" in the equation) into the predicted distribution. This is achieved by concatenating (joining) the kids' painting with the noisy image at each step (xt) within the denoising neural network (ϵθ). This ensures that the model considers the underlying structure and shapes present in the original kids' painting while adding color.

3.2.3. Combining Both Conditioning Techniques:

By combining both these conditioning methods during the inference process, Equation (1) is transformed into Equation (2). This new equation essentially represents a joint probability distribution conditioned on both the kids' painting (l) and the color scribbles (s). This means the model considers both the initial image and the user's color guidance when creating the final output.

3.2.4. Training the Model for Conditioned Colorization:

Given Equation (2), we train our proposed system using real-world pairs of kids' paintings and their corresponding color scribbles. We employ an L1 loss function (Equation (3)) to evaluate the effectiveness of the training process. This loss function essentially measures the difference between the actual noise level (ϵ) and the noise level predicted by our system (ϵθ) at each step, considering the kids' painting, color scribbles, and the current noise level of the image. By minimizing this difference, we train both the denoising model (ϵθ) and the encoder (gθ) to work together and achieve accurate colorization based on the user's input.

1. **Experiments**

This section delves into the practical evaluation of our proposed system for colorizing kids' paintings using diffusion models and user-provided color scribbles. Here, we explore the data preparation process, implementation details, and training regimen.

4.1. Dataset Preparation: Building the Foundation

In our study on line art colorization, we utilized a subsample of colorful illustrations from the publicly available dataset "safe Danbooru2021" [4]. To ensure only colorful images were included, we filtered out grayscale and monochrome images using relevant tags ("grayscale" and "monochrome"). The dataset was then divided into 200,000 images for training and 13,000 images for testing purposes.

For generating synthetic line art versions of the kids' paintings, we employed two methods: SketchKeras [14] and Sketch Simplification [20]. During training, the system randomly selected either the SketchKeras or Sketch Simplification method with a 50% chance for each image. Figure 3 showcases an example of the generated data.

To train the model to handle user-provided color guidance effectively, we simulated realistic color scribbles. This involved randomly sampling the number of scribbles (values chosen uniformly between 4 and 25), thickness (ranging from 1 to 4 pixels), and length (varying from 5 to 30 pixels). Additionally, to prevent bias towards white coloration due to many illustrations having white backgrounds, we filtered out any synthetic scribbles containing more than 60% white pixels.

4.2. Implementation Details: Bringing the Pieces Together

In our implementation for line art colorization, we built upon existing work, drawing inspiration from the methodology presented in [9]. To enhance computational efficiency, we incorporated self-attention and cross-attention mechanisms exclusively in the bottleneck layer of the denoising model (ϵθ).

For optimization and training, we employed the Adam optimizer with a learning rate of 2e-5 and utilized a cosine warm-up schedule for 5,000 training steps. A batch size of 40 was utilized to balance computational resources and training stability.

To extract color features effectively, we introduced an encoder (gθ) dedicated to this task. This encoder shared the same architecture as the denoising model (ϵθ) but featured only one residual block per layer to enhance efficiency. Both the denoising model and encoder were trained jointly from scratch.

Prior to training, all kids' painting and color scribble images underwent preprocessing, being resized to a common resolution of 256x256 pixels and normalized to the range of -1 to 1 to facilitate model convergence and stability.

For training purposes, we employed a powerful setup utilizing NVIDIA RTX 2080 Super GPUs. The final model underwent training for 80 epochs to ensure convergence and optimal performance.

1. **Putting Our Model to the Test**

This section evaluates the effectiveness of our colorization framework by comparing it to two leading user-guided line art colorization approaches ([3, 23]). We employ a combination of quantitative and qualitative assessments for a well-rounded analysis.

5.1. Quantitative Evaluation

To objectively compare the performance of different methods in line art colorization, we utilized three established metrics:

1. Structural Similarity Index (SSIM) [24]: This metric evaluates a model's ability to preserve structural details and shapes from the original image. Higher SSIM scores indicate better retention of these essential elements.
2. Learned Perceptual Image Patch Similarity (LPIPS) [25]: LPIPS goes beyond pixel-level comparisons by measuring perceptual similarities between two images. It assesses how visually similar the generated colorized output appears to a human observer compared to the original colored image. Lower LPIPS scores suggest higher perceptual similarity.
3. Frechet Inception Distance (FID) [8]: This metric measures the overall perceptual similarity between two sets of images, providing a broader perspective on the model's ability to generate colorizations that statistically resemble real-world color distributions.

To ensure fairness in comparison, all three methods were retrained using the same dataset and their default parameters as outlined in the respective research papers. Subsequently, we evaluated the models' performance on a separate set of 13,000 test images. The color illustrations in this set served as the ground truth, against which the generated images with color hints from each method were compared using the mentioned metrics.

Results and Analysis:

Our method demonstrates a clear advantage across multiple metrics:

* SSIM: Our approach retains 15% and 2% more structural information compared to the other two state-of-the-art methods, indicating superior preservation of details within the kids' paintings.
* LPIPS: Our method surpasses [3] and achieves comparable results to [26] in terms of perceptual similarity.
* FID: We outperform both methods on the FID metric, signifying that our model generates colorizations that are statistically closer to real-world color distributions of colored images.

The evaluation also revealed that using only implicit conditioning (without explicit inclusion of the kids' painting) led to lower performance compared to our full method, which utilizes both the painting and color scribbles. This highlights the effectiveness of our dual conditioning approach.

5.2. Qualitative Evaluation

While quantitative metrics offer valuable insights, visual assessment remains crucial in evaluating the quality of colorized images. Therefore, we conducted a qualitative evaluation where human observers analyzed the outputs from our method and the two comparison approaches.

The qualitative assessment confirmed the trends observed in the quantitative metrics. Our colorized images consistently exhibited:

* High-Quality Details: Fine details and textures present in the original kids' paintings were effectively preserved during the colorization process.
* Visually Appealing Colorization: The color choices were aesthetically pleasing, creating a visually engaging experience.
* Accurate Color Representation: Color shades were more accurately represented in the final images compared to the other two methods.

By combining evidence from both quantitative and qualitative evaluations, we can confidently conclude that our proposed system outperforms existing solutions in colorizing kids' paintings using diffusion models and user-provided color scribbles.



Figue 4: Example outcomes from baseline models and our network. We employed two CNN series ((b) Iizuka et al. [27], (c) Zhang et al. [28]) and two GAN series ((d) CycleGAN [29], (e) Pix2Pix [30]) for comparison, and two versions of AutoCaCoNet[31] (f, g) and finally (h) is our network.

1. **Conclusion**

This paper introduces a novel approach for user-guided colorization of kids' paintings, leveraging the capabilities of conditional Diffusion Models. Our method stands out by employing a combined implicit and explicit conditioning strategy.

* 1. Key Advantages of Our Approach:
     1. Robust Structural Generation: The implicit conditioning enables the model to effectively capture and preserve the intricate details and shapes present in the original kids' paintings. This ensures that the colorization process respects the artist's initial vision and enhances the underlying structure.
     2. Accurate User Color Representation: The explicit conditioning, facilitated by incorporating user-provided color scribbles, empowers users to guide the colorization process according to their artistic preferences. The model adeptly translates these color hints into a realistic and visually appealing color palette.

6.2 Validation of Success:

Extensive experimentation on a large-scale dataset yielded compelling results, demonstrating the superiority of our proposed method over existing techniques in both quantitative and qualitative measures.

* Quantitative Superiority: Metrics such as SSIM, LPIPS, and FID confirm that our approach excels at preserving structural information, achieving high perceptual similarity with ground-truth images, and generating colorizations that are statistically closer to real-world color distributions.
* Qualitative Confirmation: Visual inspection revealed that our method consistently produced images with greater detail retention, a more aesthetically pleasing color palette, and a more accurate representation of the intended colors compared to other state-of-the-art techniques.

1. **Future works**

This research paves the way for exciting possibilities in future advancements in user-guided image colorization. Exploring different conditioning techniques, incorporating additional user interaction methods, and investigating the potential for colorization beyond kids' paintings are all promising avenues for further exploration.

1. **References**

[1] Mi, J., Ma, C., Zheng, L., Zhang, M., Li, M., & Wang, M. (2023). WGAN-CL: A Wasserstein GAN with confidence loss for small-sample augmentation. Expert Systems with Applications, 233, 120943.

[2] Brizuela, N. G., Johnston, T. S., Alford, M. H., Asselin, O., Rudnick, D. L., Moum, J. N., ... & Lee, C. Y. (2023). A Vorticity‐Divergence View of Internal Wave Generation by a Fast‐Moving Tropical Cyclone: Insights From Super Typhoon Mangkhut. Journal of Geophysical Research: Oceans, 128(5), e2022JC019400.

[3] Fu, Y., Zhong, H., Cui, J., Liu, H., Huang, C., & Wang, J. (2023, August). User-Guided Anime Line Art Colorization with Spatially-adaptive Normalization. In 2023 IEEE Smart World Congress (SWC) (pp. 1-8). IEEE.

[4] Lan, Z., Maeda, K., Ogawa, T., & Haseyama, M. (2023). Multi-Label Classification in Anime Illustrations Based on Hierarchical Attribute Relationships. Sensors, 23(10), 4798.

[5] Stypułkowski, M., Vougioukas, K., He, S., Zięba, M., Petridis, S., & Pantic, M. (2024). Diffused heads: Diffusion models beat gans on talking-face generation. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (pp. 5091-5100).

[6] Wang, F., Geng, S., Zhang, D., & Zhou, M. (2024). Automatic colorization for Thangka sketch-based paintings. The Visual Computer, 40(2), 761-779.

[7] Wang, Y., Peng, X., Huang, W., Ye, X., & Jiang, M. (2023). Self‐supervised non‐rigid structure from motion with improved training of Wasserstein GANs. IET Computer Vision, 17(4), 404-414.

[8] Guo, X., & Mounjid, O. (2023). GANs training: A game and stochastic control approach. Mathematical Finance.

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

[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.

[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.

[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.

[13] Lee, S., & Park, E. (2024). AutoCaCoNet: Automatic Cartoon Colorization Network Using Self-Attention GAN, Segmentation, and Color Correction. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (pp. 403-411).

[14] lllyasviel. sketchkeras, 2017.

[15] Zare, M., Jampour, M., & Farrokhi, I. R. (2011, May). A heuristic method for gray images pseudo coloring with histogram and RGB layers. In 2011 IEEE 3rd International Conference on Communication Software and Networks (pp. 524-527). IEEE.

[16] Sýkora, D., Ben-Chen, M., Čadík, M., Whited, B., & Simmons, M. (2011, August). Textoons: practical texture mapping for hand-drawn cartoon animations. In Proceedings of the ACM SIGGRAPH/Eurographics Symposium on Non-Photorealistic Animation and Rendering (pp. 75-84).

[17] Weng, W., & Zhu, X. (2021). INet: convolutional networks for biomedical image segmentation. Ieee Access, 9, 16591-16603.

[18] Lee, Y., & Lee, S. (2020). Automatic colorization of anime style illustrations using a two-stage generator. Applied Sciences, 10(23), 8699.

[19] Ruthotto, L., & Haber, E. (2021). An introduction to deep generative modeling. GAMM‐Mitteilungen, 44(2), e202100008.

[20] Xie, F., Takagi, M., Seshimo, H., & Aono, Y. (2023). Refining Line Art from Stroke Style Disentanglement with Diffusion Models. IEEE Access.

[21] 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.

[22] 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.

[23] 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.

[24] Ding, K., Ma, K., Wang, S., & Simoncelli, E. P. (2020). Image quality assessment: Unifying structure and texture similarity. IEEE transactions on pattern analysis and machine intelligence, 44(5), 2567-2581.

[25] Mehdizadeh, M., MacNish, C., Xiao, D., Alonso-Caneiro, D., Kugelman, J., & Bennamoun, M. (2021). Deep feature loss to denoise OCT images using deep neural networks. Journal of Biomedical Optics, 26(4), 046003-046003.

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

[27] Iizuka, S., Simo-Serra, E., & Ishikawa, H. (2016). Let there be color! joint end-to-end learning of global and local image priors for automatic image colorization with simultaneous classification. *ACM Transactions on Graphics (ToG)*, *35*(4), 1-11.

[28] Zhang, R., Zhu, J. Y., Isola, P., Geng, X., Lin, A. S., Yu, T., & Efros, A. A. (2017). Real-time user-guided image colorization with learned deep priors. *arXiv preprint arXiv:1705.02999*.

[29] Zhu, J. Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision* (pp. 2223-2232).

[30] Isola, P., Zhu, J. Y., Zhou, T., & Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1125-1134).

[31] Lee, S., & Park, E. (2024). AutoCaCoNet: Automatic Cartoon Colorization Network Using Self-Attention GAN, Segmentation, and Color Correction. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (pp. 403-411).