Proceedings of the 6th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/95/20241762

AnimeSketchNet: Integrating traditional sketching and anime art via an innovative dataset

Jiyang Xie[1,3,*], Qiling Li[2,4]

[1]School of Electronic Information and Communications, Huazhong University of Science and Technology, Wuhan, China
[2]School of Energy and Power Engineering, Huazhong University of Science and Technology, Wuhan, China

[3]U202112655@hust.edu.cn
[4]1262589846@qq.com
*corresponding author

Abstract. Recent advancements in text-to-image synthesis have shown promise, yet many models struggle with generalizability and adaptability, particularly in specialized genres like anime sketches. To overcome these challenges, we introduce AnimeSketchNet, a novel framework designed to enhance model performance by effectively capturing the complexity and stylistic variability unique to anime. AnimeSketchNet leverages a diverse collection of anime sketch styles with varying levels of detail, enabling the model to learn and adapt to the distinct artistic style of anime. Our dual-branch architecture, featuring a decoupled attention mechanism, seamlessly assimilates textual descriptions along with visual features, producing high-quality anime sketch that are both detailed and faithful the reference sketch. Experimental results demonstrate that AnimeSketchNet surpasses existing state-of-the-art models in both qualitative and quantitative evaluations. This work significantly advances text-to-image synthesis by offering a powerful and flexible tool tailored specifically to the unique demands of anime art.

Keywords: anime sketch design, computer-aided design, text-to-image model, diffusion model.

1. Introduction

Sketching, one of humanity's oldest art forms, now faces challenges in the digital era, particularly in replicating the intuitive feeling of traditional drawing. While digital tools offer powerful editing and sharing capabilities, they often lack the spontaneity and fluidity which is essential for dynamic artistic exploration. Most computer-assisted sketch systems, despite their advancements, do not fully accommodate the iterative nature of sketching[1].

In recent years, extensive research has focused on generating anime and sketch styles using computer-assisted art generation [2]. However, there remains a notable gap in integrating these two styles for anime sketch generation. The anime sketch style merges the distinct aesthetics of anime with the hand-drawn qualities of sketches, creating a unique form of artistic expression that poses additional technical challenges. Traditional methods for generating anime and sketches often fail to capture the detailed nuances and dynamic features of both styles simultaneously.

DOI: 10.54254/2755-2721/95/20241762

To address these challenges, this paper introduces a curated dataset and a novel computational model designed to bridge the gap between traditional sketch techniques, modern digital needs, and the anime sketch style. The AnimeSketch1818 dataset, a key innovation of this work, was created by extracting frames from anime videos, converting them into sketches, and refining them with prompts. This diverse dataset serves as a solid foundation for training our model. Our model utilizes the diffusion model to replicate the delicate expression of traditional sketching while capturing the unique details of anime styles, offering the flexibility and efficiency of digital tools.

The contributions can be summarized as follows:

- We introduce AnimeSketchNet, a novel framework based on the frozen diffusion model, consisting of a lightweight adapter featuring decoupled cross attention and a resampler which captures fine-grained anime-style features.
- We introduce the AnimeSketch-1818 dataset, which includes a broad spectrum of anime sketch styles, ranging from simple line drawings to complex shading and texturing. This diversity supports comprehensive training and enhances the model's ability to generate high-quality anime-style images.
- Our quantitative and qualitative experiments demonstrate that AnimeSketchNet not only performs comparably to existing digital sketch tools but often exceeds them in accuracy, responsiveness, and user satisfaction, providing a powerful new tool for artists and designers.

2. Related Work

2.1. Evolution and characteristics of anime sketch style

The anime sketch style represents a unique intersection of traditional sketching and anime aesthetics, marking a novel evolution compared to conventional sketching techniques. Throughout art history, sketching styles have undergone extensive transformations—from the detailed anatomical studies of the Renaissance, through the emotive brushstrokes of Impressionism, to the abstract forms of 20th-century movements. However, the anime sketch style is relatively new, emerging as a distinct form that blends the expressive lines of sketching with the stylized features of anime, such as simplified forms, exaggerated characteristics, and dynamic expressions central to anime visual storytelling. Most existing studies focus on traditional sketches or anime-style images, highlighting the need for more targeted research in this area.

2.2. Generative models for sketch and anime style

The integration of artificial intelligence into art has evolved significantly, transforming the process of generating sketches and anime-style images. Early advancements, such as Generative Adversarial Networks (GANs) [3] and Neural Style Transfer (NST) [4], lay the foundation for AI-driven creativity. GANs, using a generator-discriminator framework, produce images that mimic reality, while NST blends the content of one image with the style of another to create novel artistic expressions. Building on these foundational technologies, more advanced models like DALL·E [5] and Imagen [6] have further pushed the boundaries by generating high-quality, complex images directly from textual descriptions. These models offer a broader generalization and can handle a wide range of styles, including the intricate demands of anime sketches. However, despite these advancements, many existing models still struggle with the generalizability required to effectively manage diverse sketching styles,including anime sketches. In the realm of anime sketch style generation, Methods like Anime2Sketch [7] focus on image-to-image style transfer, exploring anime sketch styles but not realize text-to-image generation. Additionally, works such as ImageNet-Sketch [8] and Pikachu-SDXL, do not fully solve the problem of generating anime sketch styles from textual descriptions.

DOI: 10.54254/2755-2721/95/20241762

3. Method

3.1. AnimeSketch-1818

*3.1.1. Data Collection.*While existing datasets for sketches and anime are both widely used in machine learning, there is still a lack of high-quality datasets specifically tailored for anime-style sketches. To fill this gap, we extract 4K resolution images (3840×2160) from a number of anime films and series. These images were initially annotated with descriptions generated by Blip-flan-t5-xl [9] and manually refined, creating an initial dataset of approximately 6,000 color images. We then used a modified version of Anime-Sketch-master to convert these images into an anime sketch-style dataset. This process involved segmenting the original images, applying Gaussian filtering, and merging them to achieve ultra-high-resolution processing (10240×10240). To mitigate detail loss during color-to-sketch conversion, we adjusted image parameters based on histogram analysis to enhance features, preserve texture, and reduce noise. As shown in Figure 1, under the guidance of art experts, we rigorously filtered out low-resolution and duplicate images, resulting in the final AnimeSketch-1818 dataset, which comprises 1,818 high-quality anime-style sketches. AnimeSketch-1818 is the first dataset specifically designed for text-to-image generation tasks in the anime sketch style.

*3.1.2. Data annotation.*To better align with text-to-image generation tasks, we shift from using original titles to natural language annotations, as outlined in figure 1. Initially, we experiment with BLIP and CapFilt techniques, a novel dataset curation method designed to filter out noisy titles and automatically generate image captions. However, upon manual review, we find this approach insufficient for complex or hard-to-identify content. To address this shortfall, we manually adjust each annotation to ensure precision. Specifically, to highlight the characteristics of anime sketch styles, we include relevant keywords in each image label. This detailed annotation ensures the dataset's high quality and applicability.

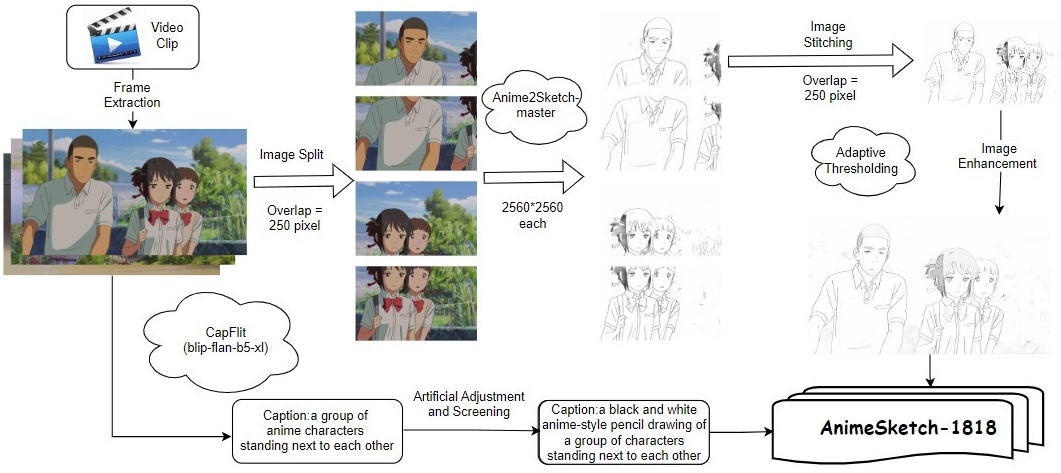

Figure 1. AnimeSketch-1818.

3.2. AnimeSketchNet

The AnimeSketchNet architecture is an innovative framework designed for anime-style sketch generation deriving from sketch. The net incorporates a resampler to capture fine-grained details and a decoupled cross-attention for image prompt.

The overall structure of AnimeSketchNet is depicted in figure 2.A key component of AnimeSketchNet is the lightweight Resampler, which stands out from traditional, more complex models by specifically aiming to minimize computational overhead while effectively processing image

DOI: 10.54254/2755-2721/95/20241762

embeddings. The Resampler achieves this through a streamlined yet potent structure that includes Linear Projection Layers and Perceiver Attention and FeedForward Network. The Linear Projection Layers transform input features into a higher-dimensional space, allowing for more refined processing. Meanwhile, the Perceiver Attention and FeedForward Network employ lightweight attention mechanisms and selectively enhance important features within the input embeddings. This integration of components ensures that AnimeSketchNet remains computationally efficient while distilling detailed features, making it especially appropriate for anime-style sketch generation.

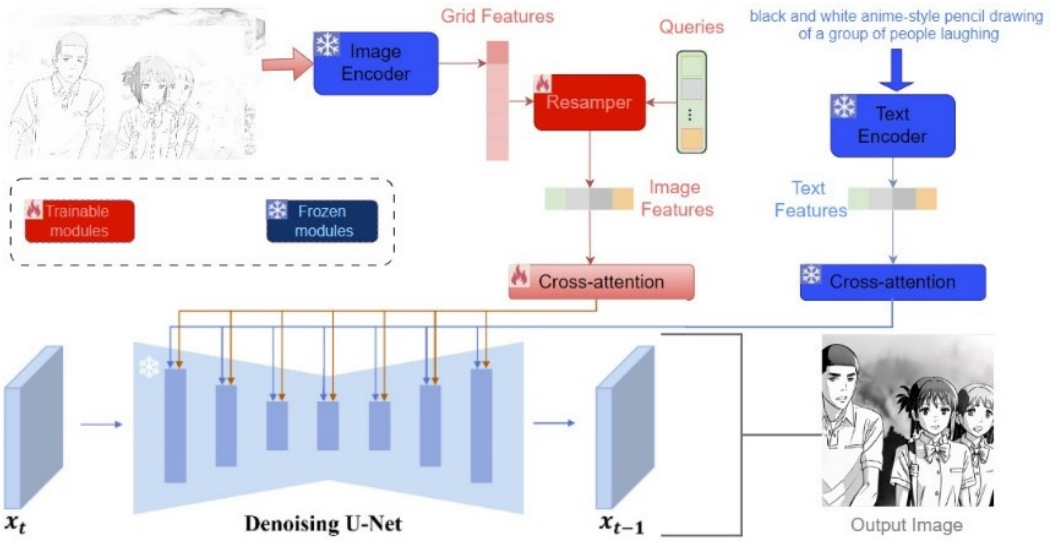

Figure 2. The overall framework of AnimeSketchNet.

4. Experiment

4.1. Implementation details

In our experiments, we utilize CLIP ViT-H/14 [10] as the image encoder and SD v1.5 [11] as the backbone network. The AnimeSketchNet framework is designed to be lightweight, allowing it to be trained on a system with a 24GB A10 GPU and 256GB of RAM. Training is carried out with the AdamW optimizer over 100,000 global steps, with a learning rate of 1×10−4 and a weight decay of 0.01. For inference, DDIM [12] is employed as the sampling acceleration algorithm, performing 50 sampling steps. The coefficient α in the dual cross-attention mechanism is set to 0.5 by default. Text inputs include keywords that identify the anime sketch style.

4.2. Quantitative evaluation

Our aim is to generate anime-style sketch images that feature diverse compositions, a wide range of elements, and visually appealing arrangements in accordance with the provided theme descriptions. To achieve this, we evaluate the text-image alignment of the generated images. For text-image alignment, we use CLIPScore based on CLIP ViT-B/32 alongside image-text matching (ITM) probability and cosine similarity between bimodal features based on BLIP-opt2.7b.

To ensure a thorough evaluation across various categories and complexities, we randomly select five prompts and compare AnimeSketchNet against several state-of-the-art models, including IP-Adapter [13], ControlNet [14], and Stable Diffusion [15]. As shown in table 1, AnimeSketchNet achieves an average CLIP score of 30.65, with a BLIP-ITM probability of 94.88%, indicating a high degree of alignment between the generated images and input text. The cosine similarity between text and image features is 0.3509 and 0.3066, reflecting strong performance compared to other models.

DOI: 10.54254/2755-2721/95/20241762

Since our model is based on IP-Adapter, we also generate 40 additional images for each prompt and compared them with IP-Adapter to assess the degree of enhancement. As shown in table 2, AnimeSketchNet outperform IP-Adapter across all metrics, particularly in BLIP-ITM probability.

Table 1. Comparative analysis of metrics for AnimeSketchNet and other methods

Method	BLIP-ITM	BLIP-Cosine	CLIP Score
AnimeSketchNet	93.95%	0.4207	30.66
ControlNet	97.68%	0.4362	31.03
IP-Adapter	71.54%	0.4292	30.97
StableDiffusion3	79.93%	0.4294	28.68

Table 2. A comparative analysis of metrics for AnimeSketchNet and IP-Adapter

Method	BLIP-ITM	BLIP-Cosine	CLIP Score
AnimeSketchNet	93.95%	0.4207	30.66
IP-Adapter	71.54%	0.4292	30.97

4.3. Qualitative analysis

In this section, we present the anime-style sketch images produced by our model. The baseline model used is SD, while other models adjust stylistic features by incorporating additional anime-style sketch images. Figure 3 demonstrates the results generalizable to different input pairs of reference images and text prompts. SD (baseline), ControlNet, and IP Adapter show instability and inaccuracies in generating anime-style sketch images. In contrast, our method accurately captures the style and texture characteristics of the images. Figure 4 reveals the robustness to varied text prompts with the same reference image. We examine the diversity of outputs generated by different methods under various prompts. Baseline, ControlNet, and IP Adapter show surprising similarity across different prompts in generation. In contrast, our method excels in both diversity and fidelity. Notably, when handling long and complex prompts, AnimeSketchNet accurately generates anime sketch style designs that are consistent with the textual descriptions.

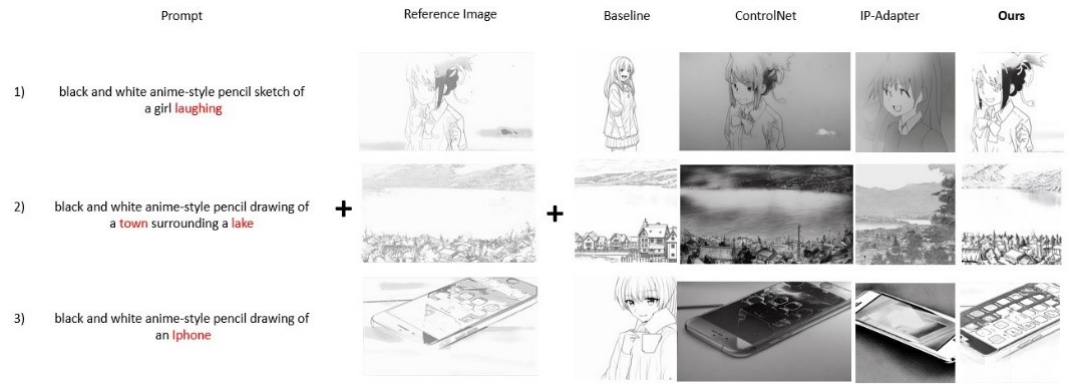

Figure 3. Results between AnimeSketchNet and other methods under specific conditions.

DOI: 10.54254/2755-2721/95/20241762

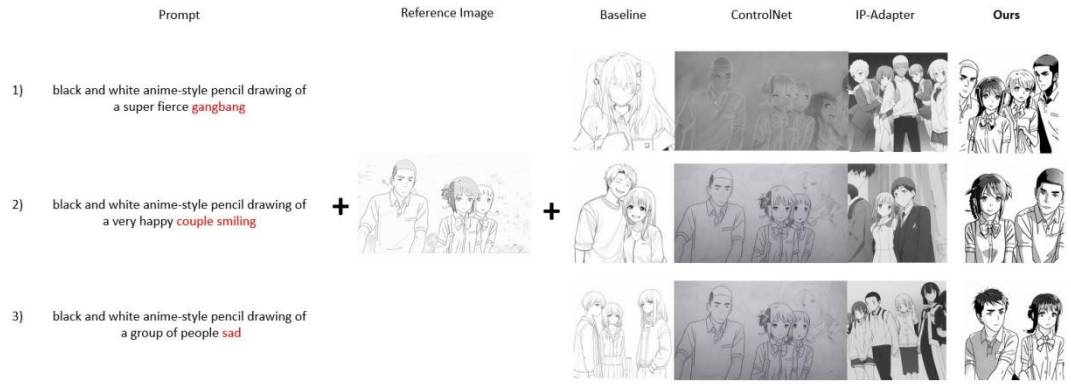

Figure 4. Results under same reference image.

5. Conclusion

In our work, we introduce AnimeSketchNet, a novel framework designed to refine the generation of anime-style sketches from line drafts with the diffusion model. We further propose dataset AnimeSketch-1818, which is meticulously created from extracted frames in various anime videos. This dataset provides a diverse collection of high-quality anime-style sketches essential for model training. By leveraging this dataset and employing a dual-branch architecture with a lightweight resampler, AnimeSketchNet demonstrates notable improvements in transforming initial anime sketches into fully realized anime images following textual prompts. This work presents a step forward in bridging the gap between traditional anime sketching and modern digital image generation, offering a valuable tool for artists and designers to refine and complete anime-style images.

References

[1] Zafer Bilda and Halime Demirkan. An insight on designers' sketching activities intraditional versus digital media. *Design studies*, 24(1):27–50, 2003.

[2] Thien Do, Van Pham, Anh Nguyen, Trung Dang, Quoc Nguyen, Bach Hoang,and Giao Nguyen. Anime sketch colorization by component-based matching using deep appearance features and graph representation. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 3154–3161. IEEE, 2021.

[3] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley,Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[4] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer usingconvolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2414–2423, 2016.

[5] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr, 2021.

[6] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479– 36494, 2022.

[7] Xiaoyu Xiang, Ding Liu, Xiao Yang, Yiheng Zhu, Xiaohui Shen, and Jan P Allebach. Adversarial open domain adaptation for sketch-to-photo synthesis. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022.

[8] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robustglobal representations by penalizing local predictive power. *Advances in Neural Information Processing Systems*, 32, 2019.

DOI: 10.54254/2755-2721/95/20241762

[9] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrappinglanguage-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

[10] Markus Hafner, Maria Katsantoni, Tino Köster, James Marks, Joyita Mukherjee, Dorothee Staiger, Jernej Ule, and Mihaela Zavolan. Clip and complementary methods. *Nature Reviews Methods Primers*, 1(1):1–23, 2021.

[11] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and BjörnOmmer. High-resolution image synthesis with latent diffusion models, 2021.

[12] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicitmodels. *arXiv preprint arXiv:2010.02502*, 2020.

[13] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.

[14] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control totext-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023.

[15] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and BjörnOmmer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.