One-Shot Reference-based Structure-Aware Image to Sketch Synthesis

Rui Yang[1,3], Honghong Yang[1*], Li Zhao[1], Qin Lei[2], Mianxiong Dong[3], Kaoru Ota[3], Xiaojun Wu[1*]

[1] Key Laboratory of Intelligent Computing and Service Technology for Folk Song, Ministry of Culture and Tourism, School of Computer Science, Shaanxi Normal University, Xi'an, 710119, China
[2]College of Computer Science, Chongqing University, Chongqing, 400044, China
[3]Muroran Institute of Technology, Muroran, Hokkaido, 0508585, Japan
{rane, yanghonghong, xjwu}@snnu.edu.cn, qinlei@cqu.edu.cn, {mx.dong, ota}@csse.muroran-it.ac.jp

Abstract

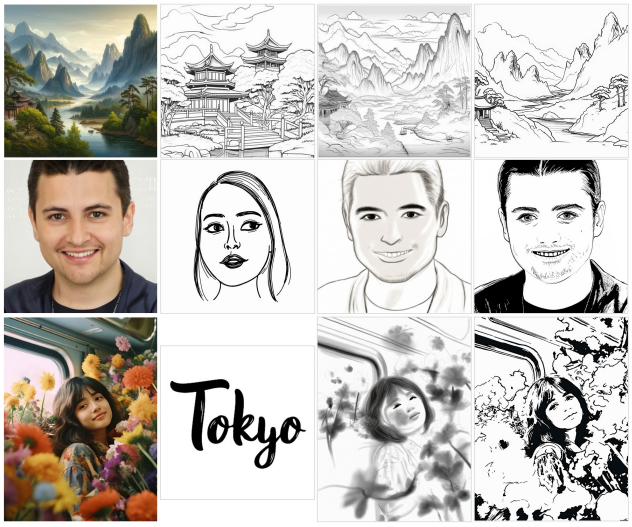
Generating sketches that accurately reflect the content of reference images presents numerous challenges. Current methods either require paired training data or fail to accommodate a wider range and diversity of sketch styles. While pre-trained diffusion models have shown strong text-based control capabilities for reference-based content sketch generation, state-of-the-art methods still struggle with reference-based sketch generation for given content. The main difficulties lie in (1) balancing content preservation with style enhancement, and (2) representing content image textures at varying levels of abstraction to approximate the reference sketch style. In this paper, we propose a method (Ref2Sketch-SA) that transforms a given content image into a sketch based on a reference sketch. The core strategies include (1) using DDIM Inversion to enhance structural consistency in the sketch generation of content images; (2) injecting noise into the input image during the denoising process to produce a sketch that retains content attributes while aligning with, yet differing in texture from, the reference. Our model demonstrates superior performance across multiple evaluation metrics, including user style preference.

Code — https://github.com/Ref2Sketch-SA

Introduction

Sketching, as a unique form of artistic expression, serves both as a powerful tool for conveying ideas and as a distinct art style. In sketch art, the brushstroke is fundamental, with varying width, direction, and pressure, contributing significantly to the depiction of form, depth, and texture. Artists use different lines to create sketches in various styles, characterized by the thickness, angle, continuity, depth, and shape of each stroke. When extracting sketches from the same scene, different artists may sketch in distinct styles. Textures in sketches are suggested through variations in line density and brushstroke manipulation, with artists using textured strokes to imply material properties and enhance depth.

Generating sketches that faithfully replicate the diversity observed in hand-drawn sketches remains a challenge. Traditional edge detection methods, such as Canny Edge Detection (Canny 1986), Holistically-Nested Edge Detection (HED) (Xie and Tu 2015), and TEED (Soria et al. 2023), primarily capture uniform coarse lines and do not emulate the artistic diversity of hand-drawn sketches. More advanced techniques like SketchKeras (Zhang 2017), PhotoSketching (Li et al. 2019), and Anime2Sketch (Xiang et al. 2022) have trained models to replicate pencil strokes using paired color images and their corresponding sketches. Recently, CLIPasso (Vinker et al. 2022), Diffsketcher (Xing et al. 2023), CLIPascene (Vinker et al. 2023) have generated vector sketches but are often limited to producing a single artistic style.



(a) Content (b) Reference (c) InstantStyle (d) Ours

Figure 1: Comparison of sketch generation results. (a) shows the original content images, and (b) presents the reference sketches. (c) displays the sketches generated by InstantStyle based on ControlNet, which, while preserving the general structure, largely loses the stylistic elements of the reference sketch. In contrast, (d) demonstrates the sketches produced by ours, which successfully aligns with the reference sketch's style while faithfully retaining the structure of the original content image.

*Corresponding Author.

To overcome the need for extensive training and to extend to arbitrary reference styles, leveraging external priors is crucial. Diffusion priors are arguably the best source for acquiring diverse visual knowledge, applied in many zero-shot settings, particularly for style transfer techniques such as StyleAligned (Hertz et al. 2024), Swap Attention Map (SAM) (Jeong et al. 2024), and StyleID (Chung, Hyun, and Heo 2024). However, these methods focus on overall style changes and cannot effectively decouple texture and contour information. These approaches face challenges in DDIM inversion (Mokady et al. 2023) operations, where fine style elements are extracted from the reference style and accurately applied to the target content image. Even though InstantStyle (Wang et al. 2024) incorporates feature sets from reference style images into specific style layers to improve the style transfer process and use additional spatial constraints (such as ControlNet (Zhang, Rao, and Agrawala 2023)) to maintain the overall structure, it tends to lose many stylistic elements of the reference sketch from the reference image when generating sketches, as illustrated in Fig. 1(c).

In this paper, we propose one-shot Reference-based Structure-Aware image to Sketch synthesis (Ref2Sketch-SA), a method for reference-based sketch generation to address these challenges. We argue that sketch generation based on the reference style of a given content image is a unique form of style transfer. Unlike previous methods, our approach mimics how an artist conceptualizes a rough outline before sketching and applies this concept to the initial latent space of stable diffusion. We introduce DDIM Inversion (Mokady et al. 2023) to enhance detail awareness in content images.

We observe that different sketch styles are not only reflected in the line style but also in varying degrees of rendering the original content image texture. Therefore, we introduce a high/low frequency filter during the denoising process, which enhances the binary representation of textures in the content image, ensuring that the resulting latent noise image reflects the importance of the texture. To prevent additional interference from the high/low frequency binary information (such as blurring or misalignment), target attributes are set as constraints during the denoising process, regularizing and/or correcting the predicted noise image in terms of image clarity and noise distribution, thereby guiding the sampling process to adjust the output. We then use Multi-Granular Edge Detection (MuGE) (Zhou et al. 2024) to extract edge details at different levels from the content image. To minimize the influence of style information in the edge map, we inject edge features into specific attention layers sensitive to spatial layout. This approach enables the creation of sketches ranging from simple object outlines to complex edge mappings with richer details. Finally, we use a style adapter to capture the line and stroke styles of the reference sketch, as well as the overall style, injecting these into specific style-sensitive layers to ensure that the generated sketch remains faithful to both the content and reference style, as shown in Fig. 1 (d).

Our approach seamlessly integrates brushstroke elements into the initial outline image, effectively merging the brushstroke style of the reference sketch with the texture and outline of the original content image. Leveraging the powerful generative capabilities of large models and the style extraction abilities of the IP-Adapter, we focus on guiding the model to generate more accurate reference-style sketches while better preserving the accuracy of the content image. Furthermore, our solution enables flexible application through controlled granularity of edge and texture injection, creating novel styles that have not been seen before. Fig. 2 shows desirable attributes of using IP-Adapter and inversion for sketch generation. Extensive experiments conducted on various standard benchmarks demonstrate our superiority in terms of sketch quality, flexibility, and applicability.

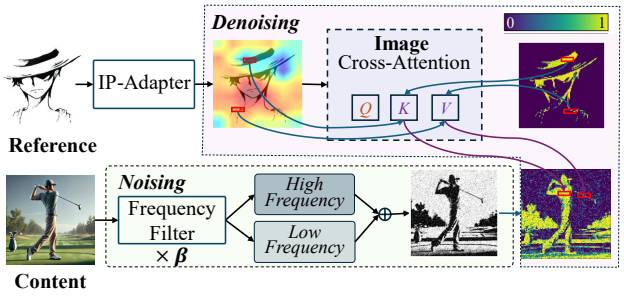


Figure 2: Ref2Sketch-SA integrates high and low-frequency information during the noising process to retain the contours and textures of the content image in the initial latent noisy sketch image. The self-attention (SA) and cross-attention mechanisms in the denoising process are crucial for reference-based structure-aware sketch generation. The input reference sketch, processed through IP-Adapter Embeds, captures global image features that, after cross-attention, yield similar attention scores. By scaling the K and V embeddings, the reference sketch's features—such as texture depth, brushstroke style, and line characteristics—are seamlessly applied during the denoising process of the content image. The integration of targeted regularization ensures the high quality of the final output.

Our contributions can be summarized as follows:

1. The proposed Ref2Sketch-SA demonstrates the ability to generate sketches faithful to both the reference sketch style and the layout of the content image through an adjustment-free strategy.
2. The design of noise and denoising correction, enhanced by content texture, allows the final sketch to fully render the reference sketch style, including both line and texture styles.
3. By injecting features into different attention layers, the generated sketch retains the reference style while avoiding content leakage from the reference image.

Related Work

Sketch Generation

Sketch generation, an advanced form of edge detection, has evolved significantly from traditional methods to deep learning-based techniques. Initially, edge detection focused

on capturing abrupt changes in image intensity, with Canny's edge detector (Canny 1986) being a foundational tool. The advent of deep learning led to more sophisticated models, such as HED (Xie and Tu 2015) and BDCN (He et al. 2019), which improved edge detection by learning richer representations and enhancing perceptual quality. Recent methods like TEED (Soria et al. 2023), UAED (Zhou et al. 2023), and DexiNed (Poma, Riba, and Sappa 2020) have further refined accuracy and detail.

However, these methods often miss the artistic subtleties required for full sketch extraction, including the style and texture of sketch lines. This prompted a shift towards models that can capture artistic nuances, transitioning from pure edge detection to more comprehensive sketch extraction. Techniques like SketchKeras (Zhang 2017), Anime2Sketch (Xiang et al. 2022), and APDrawingGAN (Yi et al. 2019) mimic artistic brushstrokes but rely heavily on paired datasets, limiting their adaptability. CLIPasso (Vinker et al. 2022) and similar methods offer greater versatility by adapting to various artistic styles without direct style pairing.

Recent approaches have focused on learning from unpaired data or single reference sketches to lower the barrier for personalized sketch models. However, these methods face challenges in maintaining style consistency and transferability, particularly with diverse and unseen reference styles. Methods like Ref2Sketch (Ashtari et al. 2022) and Semi-Ref2Sketch (Seo, Ashtari, and Noh 2023) attempt to address these issues through style clustering but still struggle with the nuances of different artistic styles. More recently, MixSA (Yang, Wu, and He 2024) has been the first reference-based training-free sketch generation method, though it emphasizes stroke-level details at the expense of global style features.

Image Stylization

Generating sketches that align with both the content layout and the reference style is a specialized form of image style transfer. The advent of diffusion models has significantly advanced style transfer techniques, with self-attention and cross-attention blocks playing a crucial role in defining the spatial layout and content of generated images. Techniques like Prompt-to-Prompt (Hertz et al. 2022) and P+ (Voynov et al. 2023) demonstrate how attention mechanisms can preserve structure while allowing greater control over style and semantics during synthesis. CIA (Alaluf et al. 2024) uses cross-image attention to establish implicit semantic correspondences, while Swap Self-Attention (Jeong et al. 2024) shows that style elements are best reflected during upsampling, though content leakage can occur in bottleneck and downsampling stages.

IP-Adapter also excels at style transfer, but its reliance on dual cross-attention for text and images can weaken text control and lead to content leakage. InstantStyle (Wang et al. 2024) mitigates this by subtracting features within the same feature space, though it, like other methods, often employs additional spatial constraints (e.g., ControlNet (Zhang, Rao, and Agrawala 2023)) for image-to-image generation. However, integrating these constraints can dilute style intensity, causing a trade-off between spatial accuracy and stylistic fidelity.

Recent methods like RB-Modulation (Rout et al. 2024) and DEADiff (Qi et al. 2024) offer solutions by seamlessly combining content and style or by decomposing layers to align semantics with coarse details and style with finer details. B-LoRA (Frenkel et al. 2024) further refines this by learning weights that separate style and content within specific blocks. Our work builds on these insights, focusing on identifying and optimizing the layers most crucial for style transfer to effectively separate and control style and content.

Methodology

Our objective is to generate an output sketch S^o that retains the structural details of a given content image C^i while adopting the stylistic attributes of a reference sketch R^i. As illustrated in Fig. 3, our approach leverages a pretrained text-to-image diffusion model, specifically SDXL (Podell et al. 2023), which is an advanced iteration of Stable Diffusion (Rombach et al. 2022). The method includes a series of steps: binary filtration, latent noising, texture enhancement, and the integration of style and layout into the diffusion process, all of which contribute to the final output.

Preliminary Observations

Diffusion models have emerged as powerful tools for stylized image generation, particularly when ensuring style consistency through textual references. However, direct image-to-image stylization remains less explored. Techniques like ControlNet (Zhang, Rao, and Agrawala 2023) employ spatial constraints to maintain content fidelity, but these can disrupt the natural brushstrokes and styles in sketch generation, leading to overly rigid lines. A simpler approach involves using sparse edges, which focus on key structural details while maintaining content integrity and promoting textural expression.

Image inversion techniques, such as SDEdit (Meng et al. 2021), FreeEnhance (Luo et al. 2024), and StyleID (Chung, Hyun, and Heo 2024), offer ways to preserve spatial structure while incorporating textural details. However, these techniques alone often fall short, necessitating our more integrated approach.

Texture Enhanced Module

In our method, the first step involves binary filtration, where the content image C^i is processed to emphasize structural details. This is done by applying the Otsu method to calculate an optimal threshold k^*, which maximizes between-class variance:

$$\sigma_B^2(\beta{\times}k) = \omega_0(\beta{\times}k)\omega_1(\beta{\times}k)\left[\mu_0(\beta \times k) - \mu_1(\beta \times k)\right]^2, \tag{1}$$

where β controls the sparsity of the binary image. This filtration allows flexible adjustment of content emphasis, ensuring that key structural details are retained in the subsequent processing stages.

Following binary filtration, the diffusion model synthesizes images by progressively denoising an initially noisy image, effectively reversing the noise addition process. For

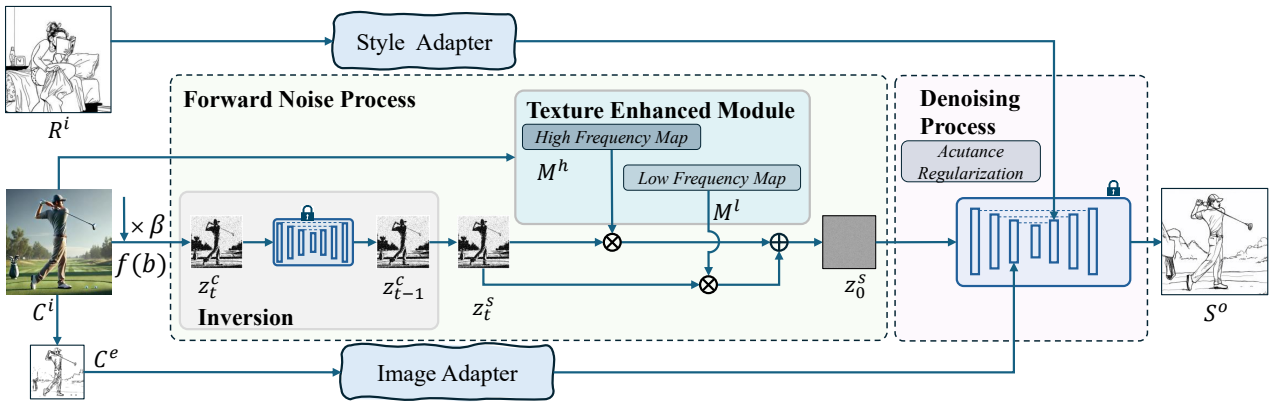


Figure 3: An overview of our Ref2Sketch-SA framework. The process begins with an input content image C^i and a reference sketch R^i. The content image C^i undergoes a binary transformation using the function $f(b)$ to generate a binary image, followed by image inversion to produce a structurally consistent image z_t^c. Meanwhile, the texture-enhanced module generates a high-frequency map M^h and a low-frequency map M^l from C^i, which are used to adaptively blend noise into the image, resulting in the noisy image z_0^s. The style adapter injects the stylistic features of R^i into the process. During the denoising process, acutance regularization is applied to guide the final sketch S^o generation, ensuring the output is faithful to both the reference style and the content structure. The edge image C^e is also utilized to maintain structural details during the sketch synthesis.

each timestep $t \in \{T, T-1, \ldots, 1\}$, noise ϵ_t is added to C^i, producing:

$$z_t^c = \alpha_t C^i + \sigma_t \epsilon_t, \tag{2}$$

where α_t and σ_t are parameters defined by the noise schedule. The network ϵ_θ then estimates ϵ_t by minimizing:

$$\arg\min_\theta \mathbb{E}_t \|\epsilon_t - \epsilon_\theta(z_t^c; t, y)\|^2, \tag{3}$$

where y is an optional conditioning signal. The image generation process alternates between noise estimation and image updating, starting from $z_T^c \sim \mathcal{N}(0, I)$ and utilizing the DDIM sampling algorithm (Mokady et al. 2023). To balance consistency and creativity, methods like SDEdit and SDXL adjust the noise level with a hyper-parameter t_0, where a smaller t_0 leads to more content-consistent but less detailed images.

To further refine the structural representation, we generate high-frequency (M^h) and low-frequency (M^l) maps from the filtered image. The high-frequency map emphasizes edges and fine details, while the low-frequency map captures broader content features. These frequency maps are used to selectively blend the noise-added images, ensuring that both detailed and general structures are preserved in the output.

Given that sketch generation requires binary expression of high and low regions rather than the original image, we apply lighter noise to high-frequency regions, such as edges and corners, to protect their original patterns. Conversely, low-frequency regions are exposed to stronger noise, promoting detail generation and refinement. We add noise corresponding to timestep t to the original image, and the diffusion model iteratively denoises until reaching timestep $t = 0$. During this process, gradient-guided sampling uses guidance generated by a predefined energy function:

$$\epsilon_t = g(x_t; y) + \alpha_t \nabla_y g(x_t; y), \tag{4}$$

and adapts the weighted average of two noise images based on the image frequency in each region:

$$z_t^{c,\text{low}} = M_l \left(z_t^{c,\text{low}} + \sigma_l \epsilon_t^{\text{low}} \right), \tag{5}$$

where M_l is the mask for low-frequency regions.

To prevent the distribution of the weighted average noise image from violating the assumed prior distribution of the diffusion process $\mathcal{N}(\alpha_t^2, \sigma_t^2)$, we use a scaling factor (shown in the attached formula) to readjust the noise image, calibrating the distribution.

Layout and Reference Injection

The IP-Adapter (Ye et al. 2023) is highly effective at extracting style while preserving content structure. Building on this, we extend the InstantStyle (Wang et al. 2024) framework, which uses cross-attention mechanisms to decouple content and style, into our approach. Our method involves selectively injecting reference style features into specific style blocks and using an edge map with subtracted style information to extract structural content. This process ensures consistency without content leakage and eliminates the need for cumbersome weight tuning.

We specifically target key layers within the diffusion model that handle style and spatial layout separately. For instance, we focus on up blocks.0.attentions.1 and down blocks.2.attentions.1, which capture style elements (such as color, material, and atmosphere) and spatial layout (structure and composition) respectively. This targeted approach can be expressed by the decoupled cross-attention strategy:

$$Z_{\text{new}} = \text{Attention}(Q, K^t, V^t) + \lambda \cdot \text{Attention}(Q, K^i, V^i), \tag{6}$$

where Q, K^t, V^t are the query, key, and value matrices for the text attention operation, and K^i and V^i are the keys and values for the image. Given the query features Z and image features c_i, $Q = ZW_q$, $K^i = c_iW_k^i$, $V^i = c_iW_v^i$.

By injecting image features into these specific attention layers, we achieve a seamless style transfer that prevents content leakage while maintaining the integrity and strength of the style. This mechanism also reduces the number of parameters required by the adapter, enhancing text control ability and making it applicable to other attention-based feature injections for tasks like editing.

Denoising with Regularization

To ensure the generated sketch closely resembles the target while preserving sharpness, we employ gradient-guided sampling:

$$\hat{\epsilon_t} = \epsilon_\theta(z_t^c; t, y) + \lambda \nabla_{z_t^c} g(z_t^c; t, y), \tag{7}$$

where $g(z_t^c; t, y)$ measures deviation from the sketch, and λ controls the influence of this guidance.

Additionally, we apply Acutance Regularization to enhance the perceived sharpness and detail in the generated sketches. This is achieved by using a Sobel operator to compute the gradient magnitudes of the image. The Sobel operator is a discrete differentiation operator that calculates an approximation of the gradient of the image intensity function. It effectively highlights areas of high spatial frequency, which correspond to edges in the image, thus enhancing the clarity of line work in the sketch. The gradient magnitude is computed as follows:

$$G_x = \begin{bmatrix} -1 & 0 & +1 \\ -2 & 0 & +2 \\ -1 & 0 & +1 \end{bmatrix} * I \quad \text{and} \quad G_y = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ +1 & +2 & +1 \end{bmatrix} * I, \tag{8}$$

where G_x and G_y represent the gradients in the horizontal and vertical directions, respectively, and I is the image. The overall gradient magnitude is then computed as:

$$G = \sqrt{G_x^2 + G_y^2}. \tag{9}$$

This gradient magnitude G highlights the edges and fine details in the image, which are crucial for the fidelity of the sketch. The Acutance Regularization loss is defined as:

$$\mathcal{L}_{acu} = -\frac{1}{HW} \sum_{i=0}^{H-1} \sum_{j=0}^{W-1} G(i, j), \tag{10}$$

where H and W are the height and width of the image, respectively, and $G(i, j)$ is the gradient magnitude at pixel (i, j). By minimizing this loss, we encourage the model to produce sketches that emphasize sharp edges and clear structural details, thereby ensuring that the sketches retain high fidelity to the original content while exhibiting clear, artistically enhanced styles.

Regularizing the Denoising. With the help of these regularizations, we additionally insert a revising step at the end of updating in each denoising iteration. Specifically, the sampling result z_{t-1} in each denoising operation is altered by $\hat{z}_{t-1}$:

$$\hat{z}_{t-1} = z_{t-1} - \zeta \nabla_{z_t} \mathcal{L}_{acu}, \tag{11}$$

where ζ is set to 3.

Experiments

We implemented our method on Stable Diffusion XL (SDXL)[1], with all experiments conducted using a pretrained IP-Adapter[2] without further fine-tuning. We evaluated our approach through qualitative and quantitative comparisons across various styles, including general photographs, color anime, and portraits.

Metrics: Evaluating stylized synthesis is challenging due to the subjective nature of style, making simple metrics inadequate. We used a two-step evaluation process: first, applying established metrics from previous works, and second, conducting a human evaluation. Given the importance of human judgment in assessing reference-aligned sketch generation, a user study was conducted to measure the alignment of both reference style and content structure.

While traditional metrics like PSNR (Wang et al. 2004), LPIPS (Zhang et al. 2018), FID (Grigorescu, Petkov, and Westenberg 2003), and SSIM (Zhang et al. 2018) are useful, they mainly assess pixel or perceptual similarity. Our method, Ref2Sketch-SA, introduces sparse representations of the original content based on reference style, which these metrics don't fully capture. Therefore, user studies played a key role in evaluating the effectiveness of our model. We also used well-defined style datasets (e.g., Anime (Kang 2018) and 4SKST (Seo, Ashtari, and Noh 2023)) to further assess robustness.

Datasets: Four datasets were used for evaluation: 4SKST (Seo, Ashtari, and Noh 2023), FS2K (Fan et al. 2022), Anime (Kang 2018), and APDrawings (Yi et al. 2019). Details are provided in Appendix A.1.

Implementation details: All experiments were conducted on a single NVIDIA 3090 GPU, using consistent hyper-parameters across tasks, and default settings for alternative methods as per their original papers. Additional details are provided in Appendix A.2.

Qualitative Analysis

Our method aims to preserve the essence of style from reference images while ensuring alignment with the content image. Fig. 4 compares our method with various training-based models (e.g., Ref2Sketch (Ashtari et al. 2022), Semi-ref2Sketch (Seo, Ashtari, and Noh 2023)) and training-free models (e.g., StyleID (Chung, Hyun, and Heo 2024), InstantStyle (Wang et al. 2024), and IP-Adapter (Ye et al. 2023)). Training-free models like InstantStyle and IP-Adapter often rely on ControlNet (Zhang, Rao, and Agrawala 2023), which can limit their ability to accurately reflect reference styles. In contrast, our method bypasses the need for ControlNet and effectively captures the unique characteristics of both style and content, generating diverse outputs that align with the input images.

As shown in Fig. 4, our method accurately replicates brushstroke styles, line thickness, and texture patterns from the reference sketch, applying them seamlessly to the content image. Additionally, our method mitigates content leakage, as seen in the first and third rows of Fig. 4, where the

[1]https://huggingface.co/stabilityai/stable-diffusion-xl-base-1.0

[2]https://huggingface.co/h94/IP-Adapter

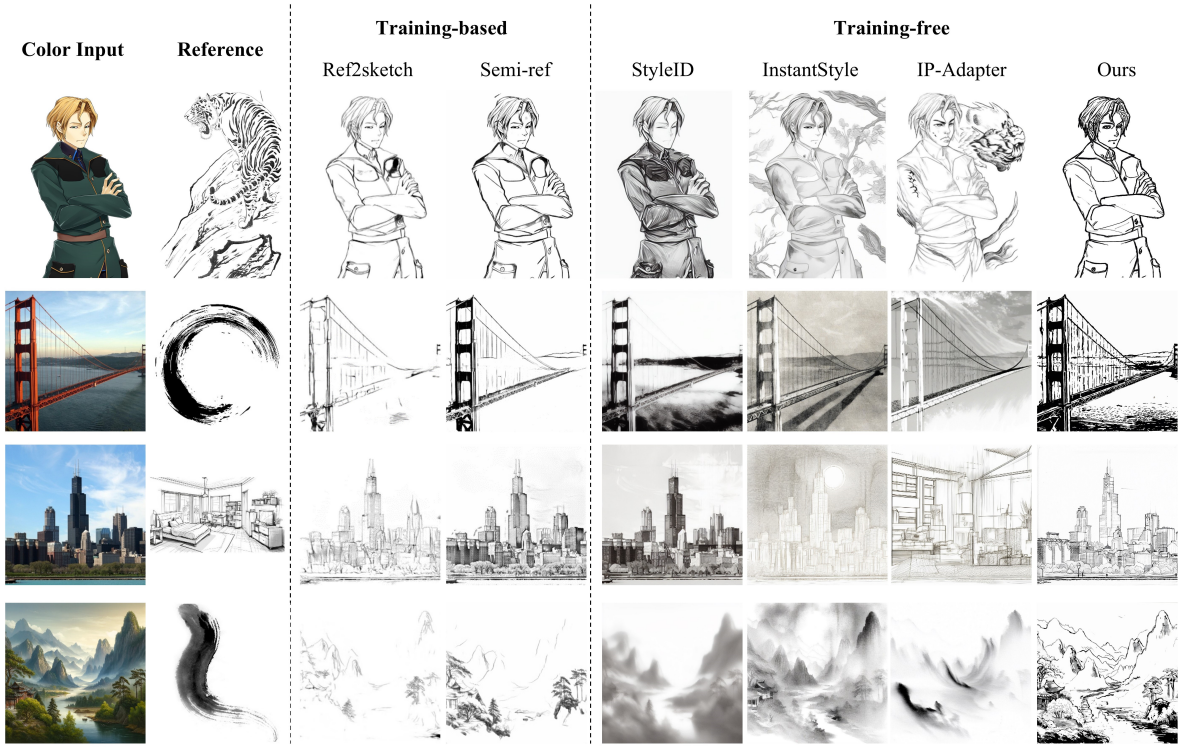


Figure 4: Comparison between training-based and training-free methods for reference-based sketch generation. The training-based methods, particularly Ref2Sketch and Semi-ref, struggle to align with reference sketch styles outside of their training datasets. While Semi-ref appears to retain the content image information in most cases, closer inspection reveals gaps in accurately aligning with the reference style. In the training-free methods, StyleID tends to average the overall tone due to its use of the AdaIN mechanism, while InstantStyle and IP-Adapter, both incorporating ControlNet, diminish the subtle stylistic nuances in the sketches. Our proposed method successfully balances both style and content, achieving superior alignment with the reference sketch style while preserving the content structure.

tiger and house are incorrectly captured by InstantStyle and IP-Adapter, but not by our approach. Notably, all prompts were set to empty to avoid interference with the generated sketches.

Compared to training-based methods like Ref2Sketch and Semi-ref2Sketch, which require clustered paired datasets, our method operates without training, delivering competitive or even superior results. For example, our approach effectively captures edges and textures, as demonstrated in the last row of Fig. 4, where training-based methods struggle with styles absent from their training data.

Fig. 5 further illustrates visual examples on the AP-Drawings test dataset, comparing our method with Infor-drawing (Chan, Durand, and Isola 2022) and other models (APDrawingGan and UPDG (Yi et al. 2020)), which require extensive data training. The results show that our training-free method achieves quality comparable to the best-performing Infor-drawing, with the flexibility to adjust parameters for any single image without the need for additional data collection or retraining.

Methods	User Score
Ref2sketch (Ashtari et al. 2022)	5.52%
Semi-ref (Seo, Ashtari, and Noh 2023)	12.95%
StyleID (Chung, Hyun, and Heo 2024)	12.10%
InstantStyle (Wang et al. 2024)	10.42%
Ours	**59.01%**

Table 1: User study results: Preference scores for different methods.

User Study

We conducted a user study with 270 participants to assess preferences regarding reference sketch fidelity and overall artistic quality. Participants evaluated sketches generated by our model and state-of-the-art models from 100 colored images using three reference sketches. Each participant was shown 20 randomly selected images and asked to choose the sketch that best matched the reference style. As shown in Table 1, our model was overwhelmingly preferred, receiving 59.01% of the votes, significantly outperforming the other methods. For more details, please refer to Appendix B.2.

(a) Content (b) APDrawing (c) UPDG (d) Infor-drawing (e) Ours

Figure 5: Visual comparison on the APDrawings test dataset. (a) Original content images; (b), (c), and (d) show results from APDrawingGan, UPDG, and Infor-drawing, all of which require extensive data training. (e) Our method generates sketches that closely approximate the high-quality results of Infor-drawing, with the added advantage of flexible parameter adjustments, eliminating the need for additional data collection or retraining.

Methods	LPIPS↓	PSNR↑	FID↓
Ref2sketch	0.2192	35.02	115.96
Semi-ref2sketch	<u>0.1271</u>	<u>35.58</u>	<u>82.18</u>
Ours	**0.1058**	**36.14**	**81.57**
SketchKera	0.2112	34.99	97.95
IrwGAN	0.1633	34.51	95.28
SwappingAutoencoder	0.2745	35.04	174.12
MUNIT	0.2582	34.23	144.82
Infor-draw	0.2130	35.05	146.86

Table 2: Quantitative results of comparison with baselines on the 4SKST dataset. The best score is annotated in bold, while the second-best scores are underlined.

Quantitative Analysis

Evaluating sketch style consistency is particularly challenging due to the lack of appropriate metrics tailored for fine-grained style coherence, especially when using nonillustrative sketch styles, as shown in Figures 1 and 5. Consequently, we perform quantitative analysis on single-style datasets, as referenced in Tables 2 and 3.

Table 2 presents the performance comparison on the 4SKST dataset (Seo, Ashtari, and Noh 2023). Notably, *Ref2sketch* and *Semi-ref2sketch* are specifically designed for sketch extraction tasks with reference sketch inputs. In contrast, methods such as IrwGAN (Xie et al. 2021), SwappingAutoencoder (Park et al. 2020), MUNIT (Huang et al. 2018), and Infor-draw (Chan, Durand, and Isola 2022) are trained for sketch generation under specific style settings. Our method achieves the best scores across all metrics (LPIPS, PSNR, and FID), showcasing its superior ability to maintain both style fidelity and content structure. Similarly,

Methods	LPIPS↓	PSNR↑	FID↓	SSIM↑
Ref2sketch	0.4655	11.09	220.41	0.7350
Semi-ref	0.4623	13.24	220.61	0.7848
StyleID	0.4734	12.72	251.79	0.7327
Ours	**0.3702**	**15.30**	**165.40**	**0.8217**

Table 3: Quantitative results of comparison with baselines on the Anime dataset (Kang 2018).

Methods	LPIPS↓	PSNR↑	FID↓
Ours	**0.4018**	**32.54**	**127.65**
w/o Inversion	0.5104	28.92	154.43
w/o TEM	0.4892	26.59	198.93
w/o AR	0.4750	28.43	184.85
w/o LI	0.4328	30.24	172.50

Table 4: Quantitative results of ablation study on FS2K dataset (Fan et al. 2022).

Table 3 reports results on the Anime dataset (Kang 2018), where our approach consistently outperforms all baselines, further validating its effectiveness in generating sketches that faithfully align with the reference style while preserving fine-grained image details.

Ablation Study

Given our focus on generating sketches based on a given content image, either Inversion or Layout Injection (LI) must be present to ensure content consistency. Table 4 provides quantitative results from our ablation study on the FS2K dataset (Fan et al. 2022), where the complete model configuration (including Texture Enhanced Module (TEM) and Acutance Regularization (AR)) consistently outperforms other configurations. This demonstrates that each module contributes to the overall performance, with the best results achieved when all modules are integrated. For detailed visual ablation experiments, please refer to Appendix B.3.

Conclusion

We introduced Ref2Sketch-SA, a training-free method for reference-based sketch generation, addressing key challenges such as balancing content retention with style enhancement and aligning content textures with reference sketch styles across different levels of abstraction. Our approach leverages DDIM Inversion to maintain structural consistency while the texture-enhanced module adaptively introduces noise, merging stylistic elements from the reference sketch. Additionally, acutance regularization during the denoising process ensures that the generated sketches stay true to the reference style while preserving the original content structure. However, a limitation of our approach is the manual adjustment of the parameter β, which, while offering flexibility for individual users, suggests that future work could explore automating β adjustment based on the style of the reference image. This could enhance the model's adaptability and further refine its output quality.

Acknowledgments

This research was partially supported by the National Natural Science Foundation of China (Nos. 62377034, 11872036, 62377033), the Fundamental Research Funds for the Central Universities of China (Grant No. GK202407007), the Key Laboratory of the Ministry of Culture and Tourism (No. 2023-02), the Innovation Team Project of Shaanxi Province (No. 2022TD26), and the Natural Science Foundation of Shaanxi Province (No. 2024JC-YBMS-503). Additional support was provided by JSPS KAKENHI (Grant Nos. JP22K11989 and JP24K14910) and JST, including PRESTO (Grant No. JPMJPR21P3) and ASPIRE (Grant No. JPMJAP2344), Japan.

References

Alaluf, Y.; Garibi, D.; Patashnik, O.; Averbuch-Elor, H.; and Cohen-Or, D. 2024. Cross-image attention for zero-shot appearance transfer. In *ACM SIGGRAPH 2024 Conference Papers*, 1–12.

Ashtari, A.; Seo, C. W.; Kang, C.; Cha, S.; and Noh, J. 2022. Reference Based Sketch Extraction via Attention Mechanism. *ACM Trans. Graph.*, 41(6).

Canny, J. 1986. A computational approach to edge detection. *IEEE Transactions on pattern analysis and machine intelligence*, (6): 679–698.

Chan, C.; Durand, F.; and Isola, P. 2022. Learning to generate line drawings that convey geometry and semantics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7915–7925.

Chung, J.; Hyun, S.; and Heo, J.-P. 2024. Style injection in diffusion: A training-free approach for adapting large-scale diffusion models for style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8795–8805.

Fan, D.-P.; Huang, Z.; Zheng, P.; Liu, H.; Qin, X.; and Van Gool, L. 2022. Facial-sketch synthesis: a new challenge. *Machine Intelligence Research*, 19(4): 257–287.

Frenkel, Y.; Vinker, Y.; Shamir, A.; and Cohen-Or, D. 2024. Implicit Style-Content Separation using B-LoRA. arXiv:2403.14572.

Grigorescu, C.; Petkov, N.; and Westenberg, M. A. 2003. Contour detection based on nonclassical receptive field inhibition. *IEEE Transactions on image processing*, 12(7): 729–739.

He, J.; Zhang, S.; Yang, M.; Shan, Y.; and Huang, T. 2019. Bi-directional cascade network for perceptual edge detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3828–3837.

Hertz, A.; Mokady, R.; Tenenbaum, J.; Aberman, K.; Pritch, Y.; and Cohen-Or, D. 2022. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*.

Hertz, A.; Voynov, A.; Fruchter, S.; and Cohen-Or, D. 2024. Style aligned image generation via shared attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4775–4785.

Huang, X.; Liu, M.-Y.; Belongie, S.; and Kautz, J. 2018. Multimodal unsupervised image-to-image translation. In *Proceedings of the European conference on computer vision (ECCV)*, 172–189.

Jeong, J.; Kim, J.; Choi, Y.; Lee, G.; and Uh, Y. 2024. Visual Style Prompting with Swapping Self-Attention. *arXiv preprint arXiv:2402.12974*.

Kang, T. B. 2018. Anime Sketch Colorization Pair. Accessed: 2024-05-18.

Li, M.; Lin, Z. L.; Mech, R.; Yumer, E.; and Ramanan, D. 2019. Photo-Sketching: Inferring Contour Drawings From Images. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2019, Waikoloa Village, HI, USA, January 7-11, 2019*, 1403–1412. IEEE.

Luo, Y.; Zhang, Y.; Qiu, Z.; Yao, T.; Chen, Z.; Jiang, Y.-G.; and Mei, T. 2024. FreeEnhance: Tuning-Free Image Enhancement via Content-Consistent Noising-and-Denoising Process. In *ACM Multimedia 2024*.

Meng, C.; He, Y.; Song, Y.; Song, J.; Wu, J.; Zhu, J.-Y.; and Ermon, S. 2021. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*.

Mokady, R.; Hertz, A.; Aberman, K.; Pritch, Y.; and Cohen-Or, D. 2023. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6038–6047.

Park, T.; Zhu, J.-Y.; Wang, O.; Lu, J.; Shechtman, E.; Efros, A.; and Zhang, R. 2020. Swapping autoencoder for deep image manipulation. *Advances in Neural Information Processing Systems*, 33: 7198–7211.

Podell, D.; English, Z.; Lacey, K.; Blattmann, A.; Dockhorn, T.; Müller, J.; Penna, J.; and Rombach, R. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*.

Poma, X. S.; Riba, E.; and Sappa, A. 2020. Dense extreme inception network: Towards a robust cnn model for edge detection. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 1923–1932.

Qi, T.; Fang, S.; Wu, Y.; Xie, H.; Liu, J.; Chen, L.; He, Q.; and Zhang, Y. 2024. DEADiff: An Efficient Stylization Diffusion Model with Disentangled Representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8693–8702.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Rout, L.; Chen, Y.; Ruiz, N.; Kumar, A.; Caramanis, C.; Shakkottai, S.; and Chu, W. 2024. RB-Modulation: Training-Free Personalization of Diffusion Models using Stochastic Optimal Control.

Seo, C. W.; Ashtari, A.; and Noh, J. 2023. Semi-supervised reference-based sketch extraction using a contrastive learning framework. *ACM Transactions on Graphics (TOG)*, 42(4): 1–12.

Soria, X.; Li, Y.; Rouhani, M.; and Sappa, A. D. 2023. Tiny and efficient model for the edge detection generalization. In *CVPR*, 1364–1373.

Vinker, Y.; Alaluf, Y.; Cohen-Or, D.; and Shamir, A. 2023. Clipascene: Scene sketching with different types and levels of abstraction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4146–4156.

Vinker, Y.; Pajouheshgar, E.; Bo, J. Y.; Bachmann, R. C.; Bermano, A. H.; Cohen-Or, D.; Zamir, A.; and Shamir, A. 2022. Clipasso: Semantically-aware object sketching. *ACM Transactions on Graphics (TOG)*, 41(4): 1–11.

Voynov, A.; Chu, Q.; Cohen-Or, D.; and Aberman, K. 2023. p+: Extended textual conditioning in text-to-image generation. *arXiv preprint arXiv:2303.09522*.

Wang, H.; Wang, Q.; Bai, X.; Qin, Z.; and Chen, A. 2024. InstantStyle: Free Lunch towards Style-Preserving in Text-to-Image Generation. *arXiv preprint arXiv:2404.02733*.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4): 600–612.

Xiang, X.; Liu, D.; Yang, X.; Zhu, Y.; Shen, X.; and Allebach, J. P. 2022. Adversarial open domain adaptation for sketch-to-photo synthesis. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 1434–1444.

Xie, S.; Gong, M.; Xu, Y.; and Zhang, K. 2021. Unaligned Image-to-Image Translation by Learning to Reweight. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 14174–14184.

Xie, S.; and Tu, Z. 2015. Holistically-nested edge detection. In *Proceedings of the IEEE international conference on computer vision*, 1395–1403.

Xing, X.; Wang, C.; Zhou, H.; Zhang, J.; Yu, Q.; and Xu, D. 2023. DiffSketcher: Text Guided Vector Sketch Synthesis through Latent Diffusion Models. In Oh, A.; Naumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Yang, R.; Wu, X.; and He, S. 2024. MixSA: Training-free Reference-based Sketch Extraction via Mixture-of-Self-Attention. *IEEE Transactions on Visualization and Computer Graphics*, 1–16.

Ye, H.; Zhang, J.; Liu, S.; Han, X.; and Yang, W. 2023. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*.

Yi, R.; Liu, Y.-J.; Lai, Y.-K.; and Rosin, P. L. 2019. APDrawingGAN: Generating Artistic Portrait Drawings from Face Photos with Hierarchical GANs. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR '19)*, 10743–10752.

Yi, R.; Liu, Y.-J.; Lai, Y.-K.; and Rosin, P. L. 2020. Unpaired portrait drawing generation via asymmetric cycle mapping. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8217–8225.

Zhang, L. 2017. sketchKeras: an u-net with some algorithm to take sketch from paints. GitHub repository.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3836–3847.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595.

Zhou, C.; Huang, Y.; Pu, M.; Guan, Q.; Deng, R.; and Ling, H. 2024. MuGE: Multiple Granularity Edge Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 25952–25962.

Zhou, C.; Huang, Y.; Pu, M.; Guan, Q.; Huang, L.; and Ling, H. 2023. The treasure beneath multiple annotations: An uncertainty-aware edge detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15507–15517.