

Image Referenced Sketch Colorization Based on Animation Creation Workflow

*Dingkun Yan^{1, 4} *Xinrui Wang²
Zhuoru Li³ Suguru Saito¹ Yusuke Iwasawa² Yutaka Matsuo² Jiaxian Guo²
¹Institute of Science Tokyo ²The University of Tokyo
³Project HAT ⁴Tokyo AI research institute, Shanda Group



Figure 1. Given reference images, our proposed method automatically synthesizes high-quality sketch colorization results that loyally match the reference color distribution and are free from artifacts.

Abstract

Sketch colorization plays an important role in animation and digital illustration production tasks. However, existing methods still meet problems in that text-guided methods fail to provide accurate color and style reference, hint-guided methods still involve manual operation, and image-referenced methods are prone to cause artifacts. To address these limitations, we propose a diffusion-based framework inspired by real-world animation production workflows. Our approach leverages the sketch as the spatial guidance and an RGB image as the color reference, and separately extracts foreground and background from the reference image with spatial masks. Particularly, we introduce a split cross-attention mechanism with LoRA (Low-Rank Adaptation) modules. They are trained separately with fore-

ground and background regions to control the corresponding embeddings for keys and values in cross-attention. This design allows the diffusion model to integrate information from foreground and background independently, preventing interference and eliminating the spatial artifacts. During inference, we design switchable inference modes for diverse use scenarios by changing modules activated in the framework. Extensive qualitative and quantitative experiments, along with user studies, demonstrate our advantages over existing methods in generating high-quality artifact-free results with geometric mismatched references. Ablation studies further confirm the effectiveness of each component. Codes are available at <https://github.com/tellurion-kanata/colorizeDiffusion>.

*Represent equal contribution to this work

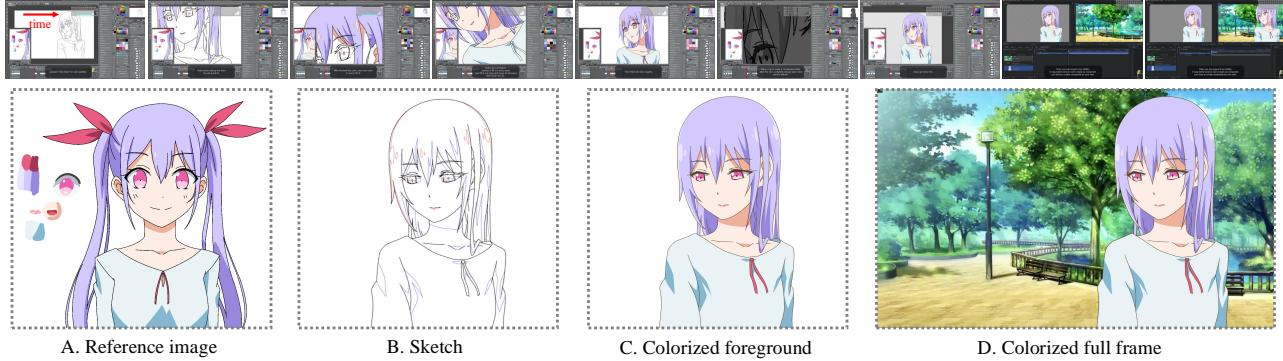


Figure 3. Illustration of colorization workflow in professional animation studios. A: character designers design characters as references. B: Senior animators draw the sketches for the key frames. C: animators colorize the figures in the sketches according to the character designs, and D: animators colorize the background of the sketches and merge foreground and background into finished frames.

1. Introduction

The past decades have witnessed the development of Animation as an artistic form, which has gained great popularity worldwide. The current workflow of animation creation is labor-intensive, and the growing demands of animation from the market are causing animation studios to fall short of hands, bringing severe problems to the industry.

The most manual labor-intensive procedure in animation production is sketch colorization, and animators working on colorization also take up the largest share among all employees in the animation production industry. To reduce the human labor needed and to automate the animation production, machine learning algorithms have been applied for sketch colorization [26, 27, 55, 56, 59]. However, current methods are still not optimal for real-world production pipelines: Text-based colorization methods [50, 57] fail to provide accurate guidance on the color and style information of the images. User-guided methods [5, 55] still involve manual operation in the process, making them less efficient. Image referenced methods [2, 47] can be seamlessly integrated into the current pipeline, but the spatial mismatches between reference images and sketches are causing severe artifacts and unexpected extra objects, which is termed as spatial entanglement in [47] and shown in Figure 2.

To build a sketch colorization framework that meets the requirements of the real-world animation production pipeline, we start with the observation of the manual sketch colorization workflow in real-world animation production. As is shown in Figure 3, the workflow consists of the following key steps: Firstly, character designers design the characters as references. Secondly, senior animators draw the sketches for each frame. Thirdly, animators colorize the figures in the sketches according to the character designs. Finally, animators colorize the background of the sketches and finish the whole colorized frames.

Following our observation, we designed a diffusion-based framework to mimic the sketch colorization workflow step-by-step. To be smoothly integrated into the current

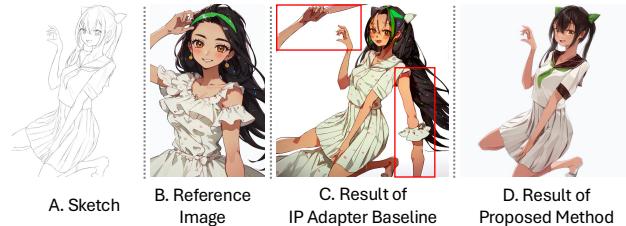


Figure 2. Illustration of spatial entanglement. We use red rectangles to highlight the spatial entangled artifacts in the result of the IP-Adapter baseline, where additional arms appear unexpectedly, and the model mistakenly synthesizes long hair.

production workflow, the proposed framework leverages a sketch image as the spatial reference and an RGB image as the color reference. Specifically, we use the high dimensional local tokens extracted by a feature extractor as reference embeddings to maintain the semantic information and adopt a multi-layer sketch encoder for precise spatial control of the background embeddings. To enable the separate colorization of foreground and background regions, we design a novel split cross-attention mechanism, where spatial masks are used to segment figures as foreground with the rest of the images as background, and corresponding LoRA weights are trained to modify the embeddings for keys and values within cross-attention layers. This design allows the diffusion model to integrate information from foreground and background independently, preventing interference and eliminating the need to adjust the well-trained backbone weights. During inference, we implement a switchable LoRA mechanism that provides precise control over the colorization process and enables different inference modes for various scenarios without changing model weights.

Experiments show that mimicking the real-world animation creation workflow yields several advantages. In qualitative analyses, our method synthesizes high-quality results loyally representing the color distribution of the reference images and free from artifacts and spatial entanglements. Quantitative comparisons also validate the superiority of the

proposed method over existing methods by common criteria and benchmarks. What’s more, user studies illustrate that candidates prefer our method subjectively.

In summary, our contributions are as follows: (1) We propose an image-referenced sketch colorization framework that can synthesize high-quality results free from artifacts and spatial entanglement by mimicking the animation production workflow. (2) We design a novel split cross-attention mechanism that enables the separate colorization of foreground and background in a single forward pass and the switchable LoRA module that allows users to switch colorization modes during inference. (3) Experiments show that our method outperforms existing methods in qualitative/quantitative comparisons and is preferred by human users in perceptive user study.

2. Related Work

2.1. Latent Diffusion Models

Diffusion Probabilistic Models [14, 41] are a class of latent variable models inspired by nonequilibrium thermodynamics [39] and have achieved great success in image synthesis and editing. Compared to Generative Adversarial Networks (GANs) [6, 7, 10, 19, 20], Diffusion Models excel at generating highly realistic images with various contexts and able to be controlled by text prompts. However, the autoregressive denoising process of diffusion models, typically computed with a U-Net [36] or a Diffusion Transformer (DiT) [3, 32], incurs substantial computational costs.

To address this limitation, Stable Diffusion (SD) [33, 35] as a class of Latent Diffusion Models (LDMs) was proposed, where a two-stage synthesis mechanism was adopted to enable diffusion/denoising process to be performed on a highly compressed latent space with a pair of pre-trained Variational Autoencoder (VAE), so as to significantly reduce computational costs. Concurrently, different researches of diffusion samplers have been conducted and proved to be effective in accelerating the denoising process [29, 30, 40, 41]. We adopt SD as our neural backbone, utilize the DPM++ solver [18, 30, 41] as the default sampler, and employ classifier-free guidance [9, 15] to strengthen the reference-based performance.

2.2. Image Prompted Diffusion Models

Currently, most diffusion models for image synthesis tasks are based on text prompts [3, 32, 33, 35]. However, there are tasks where text prompts can not provide enough information to precisely guide the image synthesis and editing, such as image-to-image translation [25], style transfer [44, 58], colorization [2, 47] and image composition [22, 51], and thus images are also used as prompts to provide reference information. The reference information extracted from prompt images varies from tasks: style transfer tasks adopt

the textures and colors from reference images, image composition tasks focus more on the object-related information, and sketch colorization requires all the above.

There are two common practices to combine image prompts with diffusion models. Given image embedding vectors extracted by pre-trained feature extraction networks, existing methods either train an adapter module to inject the reference embedding vector into the backbone [31, 48] or directly inject reference information into the backbone with attention layers [16]. However, both of them [16, 31, 48] may introduce loss or mismatch of structure when inputs are not well paired, resulting in performance deterioration. In the sketch colorization task, specifically, these adapters provide conflicting spatial information from references and lead to unacceptable artifacts, as illustrated in Figure 2.

2.3. Sketch Colorization

Sketch colorization has been a long-standing topic in computer vision. Interactive optimization-based method [42] was employed for the task, and deep-learning-based methods [2, 21, 27, 47, 56] later became the mainstream due to the ability to synthesize high-quality and high-resolution images. There are three main technical solutions for deep learning methods: text-prompted sketch colorization [21, 46, 57], user-guided sketch colorization [53, 56] and image reference sketch colorization [2, 27, 46]. User-guided methods can precisely colorize given sketches with detailed guidance from users, but the manual labor needed makes them unsuitable to be integrated into an automatic workflow. Text-prompted methods have received great popularity over recent years due to the development of Text-to-Image diffusion models, but it is challenging to precisely control colors, textures, and styles using text prompts. Image-referenced methods also benefit from the development of diffusion models, along with relevant works that enable image control [24, 31, 48, 54, 57]. However, the mismatch of reference images and sketches still results in severe deterioration. ColorizeDiffusion [47] achieved notable progress in the colorization quality, yet it still suffers from spatial entanglement, which is shown in Figure 2. In this paper, we base our method on the production pipeline of animation studios to use image references to guide colorization. We separate the foreground and background with an innovative switchable LoRA to improve colorization performance and prevent artifacts.

3. Method

Inspired by our observation of real-world animation production, we propose an image-referenced sketch colorization framework. As is shown in Figure 4, it leverages a sketch image $\mathbf{X}_s \in \mathbb{R}^{w_s \times h_s \times 1}$, a reference image $\mathbf{X}_r \in \mathbb{R}^{w_r \times h_r \times c}$ and a foreground mask $\mathbf{X}_m \in \mathbb{R}^{w_s \times h_s \times 1}$ as inputs, and returns the colorized result $\mathbf{Y} \in \mathbb{R}^{w_s \times h_s \times c}$, with

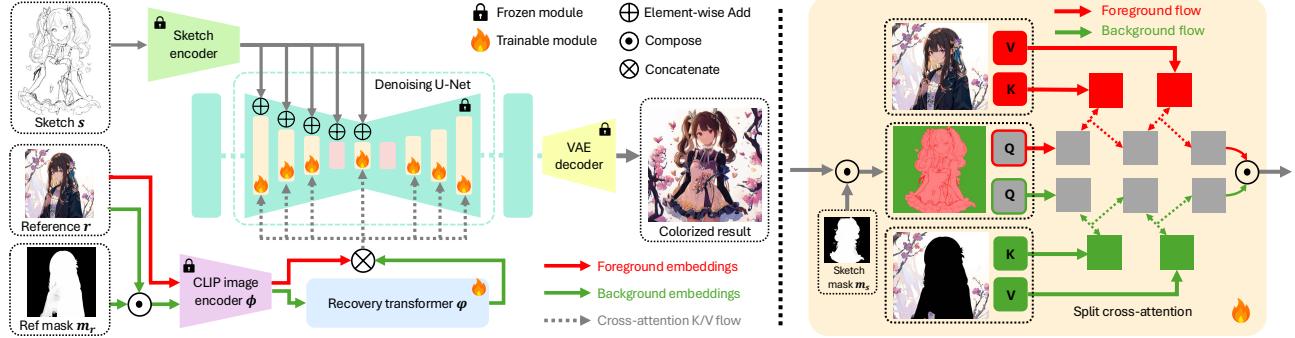


Figure 4. Illustration of the proposed framework. We use reference masks to separate reference images into foreground and background and CLIP Image encoder ϕ to extract both regions into embeddings. The background embeddings first go through the recovery transformer φ to recover detailed information, then concatenated with foreground embeddings as final K and V inputs for split cross-attention. Similar to Eq 2, the compose operation is a spatial piece-wise function employed to separate foreground and background.

w , h and c representing the width, height and channel of the images. All components of the framework are based on the animation production workflow with the following design:

Firstly, to mimic the animator’s colorizing sketches with character designs as references, we utilize a pre-trained vision transformer (ViT) to extract image embeddings as reference information. The embeddings are later injected into the diffusion backbone with split cross-attention layers. Secondly, to integrate the sketches into the framework, we adopt a multi-layer sketch encoder to inject sketch information into the latent layers of the diffusion backbone as spatial guidance. Thirdly, based on the behavior of animators separately colorizing the foreground and background of the sketch, we propose a novel split cross-attention mechanism that uses spatial masks to separate the trainable LoRA modules corresponding to the keys and values for foreground and background for training. A switchable LoRA mechanism is then applied during inference for different application scenarios with different colorization modes. More details are described in the supplementary materials.

3.1. Pretrain of the Diffusion Backbone

The backbone of the proposed framework consists of a pre-trained VAE, a sketch encoder, a denoising U-Net, and a pre-trained Vision Transformer (ViT) functioning as the image encoder from OpenCLIP-H [4, 17, 34, 37]. We denote sketch images, reference images, and ground truth as s , r , and y , respectively. The VAE encoder, U-Net, and ViT are represented by \mathcal{E} , θ , and ϕ , respectively. The timestep t starts from $T - 1$ and goes to 0, where T is the diffusion steps, set to 1000. The training objective of the diffusion model is to denoise the intermediate noisy image z_t via noise prediction:

$$\mathcal{L}(\theta) = \mathbb{E}_{\mathcal{E}(y), \epsilon, t, s, r} [\|\epsilon - \epsilon_\theta(z_t, t, s, \phi(r))\|_2^2]. \quad (1)$$

To pre-train the diffusion backbone, We initialize the VAE and U-Net with WaifuDiffusion [11] and train networks with a dynamic reference drop rate decreasing from

80% to 50% as training progresses to avoid the distribution shift mentioned by [47]. VAE, U-Net, and sketch encoder are then frozen during the training of the full framework.

3.2. Color Reference Extraction

Following the real-world animation production workflow, we utilize images as color references for sketch colorization. The commonly used ViT-based image encoder networks have two kinds of output embeddings: the CLS embeddings $E_{cls} \in \mathbb{R}^{bs \times 1 \times 1024}$ and the local embeddings $E_{local} \in \mathbb{R}^{bs \times 256 \times 1024}$, where bs represent the batch size. The CLS embeddings are projected to CLIP embedding space for image-text contrastive learning, with spatial information compressed and connected to text-level notions, and are employed as color or style references by previous image-guided methods [44, 48]. ColorizeDiffusion [47], on the contrary, reveals that local embeddings also express text-level semantics, indicating that they express more details regarding textures, strokes, and styles, enabling the network to generate better reference-based results, especially for transferring detailed textures and strokes. Therefore, the proposed method follows [47] to adopt local tokens as color reference inputs for the framework.

However, the excessive spatial information of image semantics and compositions contained in local embeddings leads to frequent occurrences of artifacts such as overflow of color regions and unexpected objects outside sketches. Illustrated in Figure 2, such artifacts widely exist in frameworks with image references [31, 44, 48]. To eliminate this problem, we follow the real-world workflow to explicitly separate the foreground and background with spatial masks during colorization and describe the detail in 3.3 and 3.4.

3.3. Split Cross-Attention

In anime images, the foreground regions and background regions differ distinctively in color distribution, color block sizes, tones, and textures. Thus, the colorization of foreground and background is separated into two independent

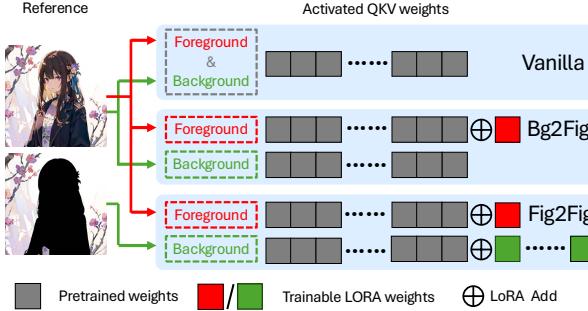


Figure 5. Based on the LoRA weights, the proposed method can merge the foreground and background features in one forward pass and switch between three inference modes. We denote the dimension of pre-trained weights as CH. The rank of foreground LoRA is fixed at 16, while the rank of background LoRA is 0.5*CH.

steps in the animation production workflow. Following this scheme, we propose a novel split cross-attention mechanism to substitute the cross-attention layers in the diffusion backbone to separately process foreground and background regions with different parameters in a single forward pass.

A split cross-attention layer consists of two groups of trainable LoRA weights W_f^t and W_b^t , which include query weights W_f^q and W_b^q , key weights W_f^k and W_b^k , and value weights W_f^v and W_b^v for foreground and background QKV projection respectively. An open-sourced animation image segmentation tool [38] is used to automatically extract the foreground mask m_s and m_r of sketches and reference images. The spatial masks can be adjusted by tuning the user defined thresholds ts_s and ts_r . Regions with pixel values larger than ts_s and ts_r are considered as foreground, otherwise background. For foreground LoRAs, we set the ranks as 16; for background LoRA, the rank is formulated as $r = 0.5 * \min(D_q, D_{kv})$, where D_q and D_{kv} are dimensions of queries and keys/values for the corresponding cross-attention layers.

As foreground and background regions of animation images feature different textures, color distribution, and color tones, injecting the embeddings of foreground and background directly into split cross-attention results in deterioration in structure preservation, synthesis quality, and stylization. Therefore, we further add a trainable recovery transformer φ to process the background embeddings and facilitate better integration of the foreground and background reference information into the diffusion backbone.

We define query inputs (forward features) as z_f, z_b , key and value inputs (reference embeddings) as e, e_b , attention outputs as y in the following sections, where the index f and b indicate foreground and background respectively. Specifically, e denotes the reference embeddings extracted from the whole reference image r , formulated as $e = \phi(r)$; and $e_b = \varphi(\phi(r_b))$, where r_b is the background region of the reference image. During training, the proposed split cross-attention can be formulated as follows:

$$y = \begin{cases} \text{Softmax}\left(\frac{(\hat{W}_f^q z_f) \cdot (\hat{W}_f^k e)}{d}\right)(\hat{W}_f^v e) & \text{if } m_s > ts_s \\ \text{Softmax}\left(\frac{(\hat{W}_b^q z_b) \cdot (\hat{W}_b^k e_b)}{d}\right)(\hat{W}_b^v e_b) & \text{if } m_s \leq ts_s \end{cases} \quad (2)$$

where $\hat{W}_f^t = W^t + W_f^t$, and W^t represents the pre-trained weights, which remain frozen during training. Similarly, \hat{W}_b^t follows the same approach.

3.4. Switchable inference mode

The application scenarios and the sketch-reference combinations in the real world may be complicated. Also, naively separating the foreground and background regions for colorization degrades the quality of background synthesis, especially when reference images have severe semantic mismatches with the sketches or have complicated backgrounds. Therefore, we design three different inference modes: *Vanilla*, *Bg2Fig*, and *Fig2Fig* for different scenarios based on the weights and reference inputs used for KV calculation. We visualize all inference modes in Figure 5.

Vanilla mode only utilizes the pre-trained weights for cross-attention modules, with W_f^t, W_b^t and recovery transformers deactivated. It's suitable for most scenarios but suffers from spatial entanglement when references are figure-only images.

Bg2Fig mode activates only the foreground-related LoRA weights W_f^t during inference and is used when reference images are figure images with complicated backgrounds. This mode outperforms *Vanilla* mode in character colorization and *Fig2Fig* mode in background generation as its foreground weights are further optimized by LoRAs.

Fig2Fig is designed for figure-to-figure colorization, where reference images are figures with simple background composition. This mode activates both LoRA weights W_f^t, W_b^t , and the recovery transformer and uses spatial masks to separate foreground/background embeddings for calculating query and key/value inputs. It effectively eliminates the spatial entanglement in reference-based sketch colorization.

4. Experiment

4.1. Implementation details

The Danbooru2021 dataset [8] is used to train and validate the proposed method. We select and process 4.8 million triples of (sketch, color, mask) images with various contents to train the framework, and use another 50,000 image triples excluded from training set for validation. Sketches were extracted by jointly using [52] and [45]. 4x H100 (94GB) GPUs are used to train the network. We at first pre-train the denoising U-Net and sketch encoder for 6 epochs with 8 hours per epoch, and then froze the backbone and trained the recovery transformer and switchable LORAs for 3 epochs with 9 hours per epoch. AdamW op-

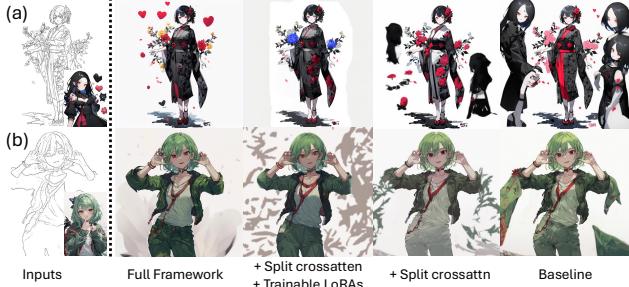


Figure 6. Results of the ablation study. The baseline model demonstrates significant spatial entanglement; incorporating split cross-attention reduces artifacts, the trainable LoRAs improve color saturation and details, and the proposed complete pipeline produces high-quality results free of artifacts. Zoom in for details.

timer [23, 28] is used to optimize the network with learning rate 0.0001 and betas (0.9, 0.999). Following [47], we dropped at least 50% of the reference inputs in all training.

4.2. Ablation study

Split cross-attention. The proposed method aims to address spatial entanglement by simulating the animation workflow. To demonstrate the effectiveness of this workflow, we set up three frameworks: 1) baseline model without split cross attention, trainable LoRAs, and recovery transformer, 2) baseline model with split cross attention but no trainable LoRAs and recovery transformer, and 3) the proposed full framework.

We show the qualitative comparison in Figure 6 to validate the effectiveness of the proposed modules. The baseline model causes severe spatial entanglement in generating additional figures in (a) and undesired clothes in (b). The application of split cross-attention mitigates the spatial entanglement but still causes artifacts and degrades the color saturation and details of the results. Collaborating split cross attention with trainable LoRAs improves the quality of results and further improves the background, but still suffers from artifacts. The proposed full framework enhanced by recovery transformers effectively eliminates spatial entanglement and synthesizes colorization results that have clear boundaries and rich details and textures, and loyally preserves the color distribution of reference images.

Inference modes. We illustrate the differences between inference modes and their use cases in Figure 7. *Fig2Fig* mode is fully mask-guided, enabling it to eliminate spatial entanglement shown in (a). However, it is less suitable for inpainting character sketches, as the region without the mask guide may suffer from the lack of reference information. *Bg2Fig* performs similarly to the *Vanilla* mode. With the help of foreground LoRA weight, the results demonstrate clearer segmentation and better stroke quality. All modes perform well for landscape sketches, which contain more structural detail than figure sketches, making them

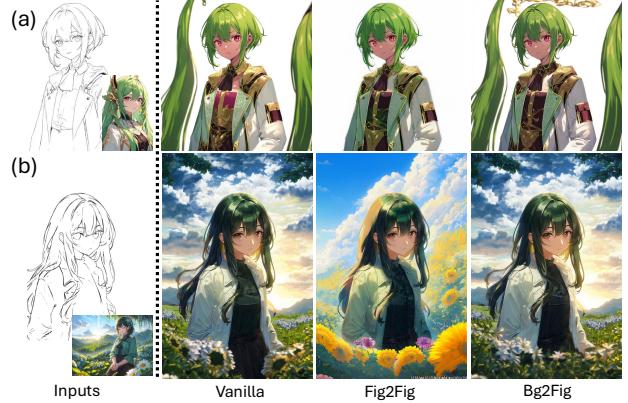


Figure 7. Colorization results with three different inference modes. *Fig2Fig* mode performs better in eliminating spatial entanglement, while *Bg2Fig* and *Vanilla* mode can generate vivid backgrounds and inpainting results.

easier to colorize.

We show the comparison of local embeddings and CLS embeddings for **color reference extraction** in the supplementary materials.

4.3. Comparison with baseline

We compare our method with six existing reference-based sketch colorization methods [2, 46, 47] to demonstrate the superiority of the proposed framework.

Baseline introduction. Two baseline methods are the combination of SD [35], ControlNet [54, 57], and IP-Adapter [48]. They adopt different cross-attention scales during denoising, labeled as *IP-Adapter* and *InstantStyle*, respectively. *IP-adapter* baseline generates results with normal cross-attention scales, while *InstantStyle* generates results using the “style transfer” weight type, which is claimed to prevent composition transfer by setting specific cross-attention scales to 0 according to [44]. Following the official document of [54], we adopted Anything v3 [49], a community-developed model, as the SD backbone for IP-Adapter-H and ControlNet_lineart_anime. All these models are officially implemented and claimed to be effective for anime-style image generation. The *T2I-Adapter* baseline simply replaces IP-Adapter with T2I-Adapter-Style [31, 43]. Specifically, we introduced quality-related prompts and textual inversion, such as “masterpiece” and “easynegative” [12], for T2I-model baselines to improve their image quality.

Qualitative comparison. We show the qualitative comparison of our proposed method and existing methods in Figure 8, where most of the existing methods suffer from spatial entanglement in various cases. In rows (a)-(d), existing methods (2)-(6) failed to distinguish foreground and background regions and, therefore, synthesized artifacts in

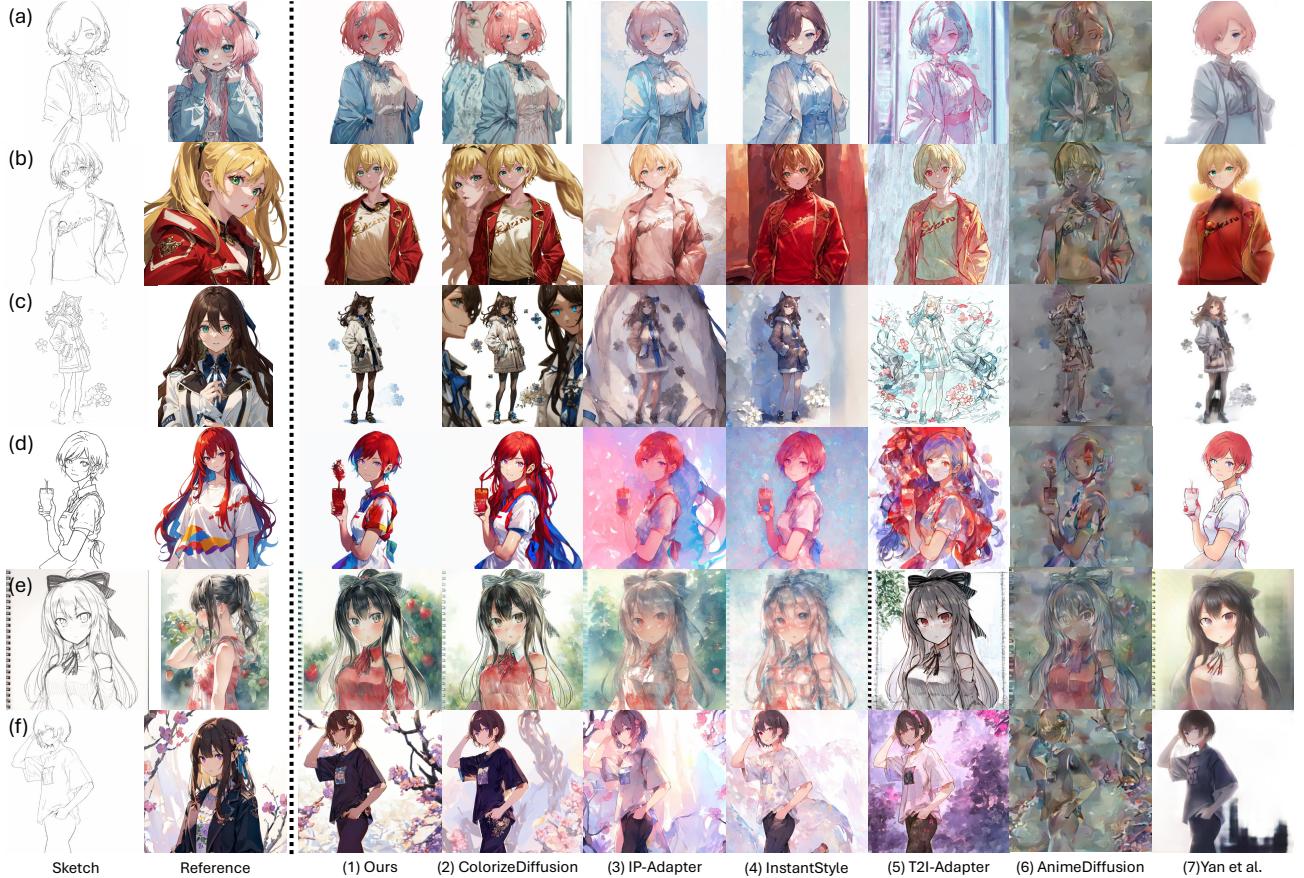


Figure 8. Qualitative comparisons between our proposed method and existing methods show that our results are visually more appealing than those of image-prompt adapters [31, 44, 48] and the GAN-based method [46]. Compared to ColorizeDiffusion [47], the proposed framework eliminates spatial entanglement and improves overall quality.

the background. Such artifacts become more obvious when reference images have complicated backgrounds in rows (e) and (f), where IP-Adapter and InstantStyle mixed reference composition with the sketch composition, making the generated results visually messy. The GAN-based method (7) successfully generated results with clear backgrounds, but the color preservation, texture transformation, and detail qualities are much poorer than diffusion-based methods.

We also show the comparison of our proposed method and adapter-based methods on semantically non-relevant sketch-reference pairs in Figure 9. Both IP-Adapter and T2I-Adapter fail to separate foreground from background and colorize the whole image with the same tone and texture. Our proposed method, on the contrary, generated results with clear region boundaries, rich texture and details, and visually pleasant bright color, with the help of the proposed spatial aware split cross-attention mechanism and switchable LoRA. The qualitative comparisons demonstrate that our proposed method is effective for arbitrary input sketch-reference pairs and achieves high-quality colorization in various use scenarios.

Quantitative comparison. Fréchet Inception Distance

(FID) [13] is a widely used quantitative criterion to evaluate the performance of image synthesis tasks. It calculates the perceptual distance of two distributions without requiring them to be semantically and spatially paired. We conduct a quantitative evaluation measured by FID on the entire validation set which includes 52k+ (sketch, reference) image pairs. Sketches are colorized with randomly selected reference images, ensuring each batch has semantically and spatially mismatched inputs.

Multi-scale structural similarity index measure (MS-SSIM), peak signal-to-noise ratio (PSNR), and CLIP score [17, 34] assess the similarity between processed images and given ground truth. It requires the guiding reference to be aligned with the ground truth when applied to image-referenced sketch colorization. To fulfill this, we selected 5000 color images as ground truth, extracted the sketches from and used thin plate spline (TPS) transformation to spatially distort them as color references to build a test set of 5000 triples of sketch, reference, and ground truth, and evaluate all the 7 methods with MS-SSIM, PSNR and CLIP score on this dataset.

We show the results of the quantitative comparison in Ta-

Table 1. Quantitative comparison between the proposed model and baseline methods. We calculated 50K FID, 5K PSNR, 5K MS-SSIM, and 5K CLIP cosine similarity of image embeddings in this experiment. †: FID evaluation randomly selected color images as references, making it close to real-application scenarios. ‡: Ground truth color images were deformed to obtain semantically paired and spatially similar references for evaluations. We also show the inference time on H100 GPU at 512*512 resolution and batch size 4.

Method	†50K-FID ↓	‡PSNR↑	‡MS-SSIM↑	‡CLIP similarity↑	Inference time(ms)
Ours	6.8327	28.9144	0.6002	0.8829	1513
ColorizeDiffusion [47]	9.5276	28.7384	0.5913	0.8775	1387
IP-Adapter [48, 49, 57]	38.9184	28.6767	0.5478	0.8672	2617
InstantStyle [44, 48, 49, 57]	40.8144	28.1090	0.4459	0.8042	2586
T2I-Adapter [31, 49, 57]	41.1569	28.1275	0.3243	0.7180	2301
AnimeDiffusion [2]	61.5999	27.8454	0.3185	0.7319	2073
Yan et al. [46]	27.0147	29.2483	0.5253	0.7634	57



Figure 9. Colorization results for semantically non-relevant sketch-reference pairs. Adapter-based methods fail to generate clear borders and high-quality textures, while our proposed method synthesizes visually satisfying results.

ble 1, where the proposed method significantly outperforms in FID, MM-SSIM, and CLIP similarity due to better texture and color synthesis. GAN-base method [46] achieved the best score in PSNR, with the proposed method ranked number 2. This is because the limited generation ability of [46] prevents it from generating complicated backgrounds and also hinders it from synthesizing bright colors and rich details of the figures. The close-to-average results make it advantageous to the calculation of PSNR [1].

User study. The quality of sketch colorization is highly subjective and easily influenced by personal preference, we thus demonstrate how different individuals evaluate the proposed method and existing methods through an user study. 40 participants are invited to select the best results with two criteria: the overall colorization quality and the preservation of geometric structure of the sketches. We prepare 25 image sets and show each participant 16 image sets for evaluation. For each image set, participants are presented the colorization results of our proposed method as well as those

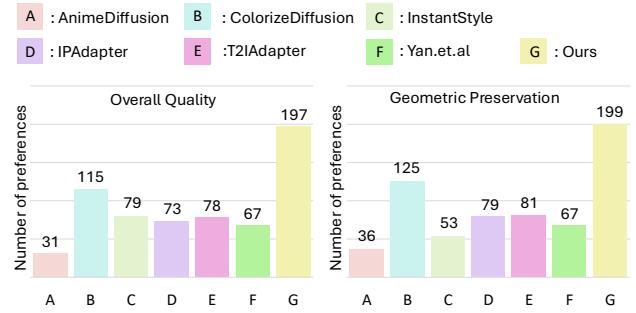


Figure 10. Results of user study. Our method is preferred across all shown methods in overall quality and geometric preservation.

generated by six existing methods.

The results of the user study are presented in Figure 10, where our proposed method has received the most numbers of preferences among all the methods presented. To further validate the comparison, we applied the Kruskal-Wallis test as a statistical method. The results clearly demonstrate that the proposed method significantly outperforms all the existing methods in terms of user preference with a significance level of $p < 0.01$. All the images shown in the user study are included in the supplementary materials.

5. Conclusion

In this paper, we proposed an image-guided sketch colorization framework inspired by real-world animation creation workflow. We implement a novel split cross-attention layer with spatial masks and corresponding LoRA weights that facilitates separate processing for foreground and background, eliminating the spatial entanglement, and also design a switchable LoRA mechanism that enables users to choose different inference modes for various use cases.

However, the proposed methods strongly depend on the quality of extracted masks, and a corresponding failure case is given in the supplementary materials. Our future work focuses on further improving the similarity with references and extending the framework for video colorization.

References

- [1] Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6228–6237, 2018. 8
- [2] Yu Cao, Xiangqiao Meng, P. Y. Mok, Tong-Yee Lee, Xuetong Liu, and Ping Li. Animediffusion: Anime diffusion colorization. *TVCG*, pages 1–14, 2024. 2, 3, 6, 8
- [3] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart- α : Fast training of diffusion transformer for photorealistic text-to-image synthesis, 2023. 3
- [4] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *CVPR*, pages 2818–2829, 2023. 4
- [5] Youngin Cho, Junsoo Lee, Soyoung Yang, Juntae Kim, Yeojeong Park, Haneol Lee, Mohammad Azam Khan, Daesik Kim, and Jaegul Choo. Guiding users to where to give color hints for efficient interactive sketch colorization via unsupervised region prioritization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1818–1827, 2023. 2
- [6] Yunjey Choi, Min-Je Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *CVPR*, pages 8789–8797. IEEE/CVF, 2018. 3
- [7] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *CVPR*, pages 8185–8194. IEEE/CVF, 2020. 3
- [8] Danbooru community, Gwern Branwen, and Anonymous. Danbooru2021: A large-scale crowdsourced and tagged anime illustration dataset. <https://gwern.net/danbooru2021>, 2022. Accessed: DATE 2022-01-21. 5
- [9] Prafulla Dhariwal and Alexander Quinn Nichol. Diffusion models beat gans on image synthesis. In *NeurIPS*, pages 8780–8794, 2021. 3
- [10] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, pages 2672–2680, 2014. 3
- [11] Reimu Hakurei. Hugging face/waifu-diffusion-v1-4. <https://huggingface.co/hakurei/waifu-diffusion-v1-4>, 2023. Accessed: DATE 2023-03-05. 4
- [12] Havoc. Easynegative. <https://civitai.com/models/7808/easynegative>, 2023. Accessed: DATE 2023-02-10. 6
- [13] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, pages 6626–6637, 2017. 7
- [14] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 3
- [15] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *CoRR*, abs/2207.12598, 2022. 3
- [16] Li Hu, Xin Gao, Peng Zhang, Ke Sun, Bang Zhang, and Liefeng Bo. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. *arXiv preprint arXiv:2311.17117*, 2023. 3
- [17] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip. https://github.com/mlfoundations/open_clip, jul 2021. 4, 7
- [18] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *NeurIPS*, 2022. 3
- [19] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, pages 4401–4410. IEEE/CVF, 2019. 3
- [20] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *CVPR*, pages 8107–8116. IEEE/CVF, 2020. 3
- [21] Hyunsu Kim, Ho Young Jhoo, Eunhyeok Park, and Sungjoo Yoo. Tag2pix: Line art colorization using text tag with secat and changing loss. In *ICCV*, pages 9055–9064. IEEE/CVF, 2019. 3
- [22] Kangyeol Kim, Sunghyun Park, Junsoo Lee, and Jaegul Choo. Reference-based image composition with sketch via structure-aware diffusion model. *arXiv preprint arXiv:2304.09748*, 2023. 3
- [23] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 5
- [24] Kohya-ss. Hugging face/controlnet-lllite. <https://huggingface.co/kohya-ss/controlnet-lllite>, 2024. Accessed: DATE 2024-01-02. 3
- [25] Gihyun Kwon and Jong Chul Ye. Diffusion-based image translation using disentangled style and content representation. In *ICLR*. OpenReview.net, 2023. 3
- [26] 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 *CVPR*, pages 5800–5809. IEEE/CVF, 2020. 2
- [27] Zekun Li, Zhengyang Geng, Zhao Kang, Wenyu Chen, and Yibo Yang. Eliminating gradient conflict in reference-based line-art colorization. In *ECCV*, pages 579–596. Springer, 2022. 2, 3
- [28] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*. OpenReview.net, 2019. 5
- [29] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ODE solver for diffusion probabilistic model sampling in around 10 steps. In *NeurIPS*, 2022. 3
- [30] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver++: Fast solver for

- guided sampling of diffusion probabilistic models. *CoRR*, abs/2211.01095, 2022. 3
- [31] Chong Mou, Xintao Wang, Liangbin Xie, Jian Zhang, Zhonggang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *CoRR*, abs/2302.08453, 2023. 3, 4, 6, 7, 8
- [32] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, pages 4172–4182. IEEE, 2023. 3
- [33] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: improving latent diffusion models for high-resolution image synthesis. *CoRR*, abs/2307.01952, 2023. 3
- [34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, volume 139, pages 8748–8763. PMLR, 2021. 4, 7
- [35] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10674–10685. IEEE/CVF, 2022. 3, 6
- [36] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, volume 9351, pages 234–241. Springer, 2015. 3
- [37] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5b: An open large-scale dataset for training next generation image-text models. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. 4
- [38] SkyTNT, infoengine1337, and not lain. anime-segmentation. <https://github.com/SkyTNT/anime-segmentation>, 2022. 5
- [39] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, volume 37, pages 2256–2265. JMLR.org, 2015. 3
- [40] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*. OpenReview.net, 2021. 3
- [41] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*. OpenReview.net, 2021. 3
- [42] Daniel Sýkora, John Dingiana, and Steven Collins. Lazy-brush: Flexible painting tool for hand-drawn cartoons. *Comput. Graph. Forum*, 28(2):599–608, 2009. 3
- [43] TencentARC. Hugging face/ip-adapter. <https://github.com/TencentARC/T2I-Adapter/tree/SD>, 2024. Accessed: DATE 2024-01-02. 6
- [44] Haofan Wang, Qixun Wang, Xu Bai, Zekui Qin, and Anthony Chen. Instantstyle: Free lunch towards style-preserving in text-to-image generation. *arXiv preprint arXiv:2404.02733*, 2024. 3, 4, 6, 7, 8
- [45] Xiaoyu Xiang, Ding Liu, Xiao Yang, Yiheng Zhu, Xiaohui Shen, and Jan P. Allebach. Adversarial open domain adaptation for sketch-to-photo synthesis. In *WACV*, pages 944–954. IEEE/CVF, 2022. 5
- [46] Dingkun Yan, Ryogo Ito, Ryo Moriai, and Suguru Saito. Two-step training: Adjustable sketch colourization via reference image and text tag. *Computer Graphics Forum*, 2023. 3, 6, 7, 8
- [47] Dingkun Yan, Liang Yuan, Yuma Nishioka, Issei Fujishiro, and Suguru Saito. Colorizeddiffusion: Adjustable sketch colorization with reference image and text. 2024. 2, 3, 4, 6, 7, 8
- [48] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *CoRR*, abs/2308.06721, 2023. 3, 4, 6, 7, 8
- [49] Yun0779. <https://civitai.com/models/9409>, 2023. Accessed: DATE 2023-06-25. 6, 8
- [50] Nir Zabari, Aharon Azulay, Alexey Gorkor, Tavi Halperin, and Ohad Fried. Diffusing colors: Image colorization with text guided diffusion. In *SIGGRAPH Asia 2023 Conference Papers*, pages 1–11, 2023. 2
- [51] Bo Zhang, Yuxuan Duan, Jun Lan, Yan Hong, Huijia Zhu, Weiqiang Wang, and Li Niu. Controlcom: Controllable image composition using diffusion model. *arXiv preprint arXiv:2308.10040*, 2023. 3
- [52] Lvmin Zhang. Sketchkeras. <https://github.com/lillyasviel/sketchKeras>, 2017. 5
- [53] Lvmin Zhang. Style2paints v5, 2023. Accessed: DATE 2023-06-25. 3
- [54] Lvmin Zhang. Controlnet-v1-1-nightly. <https://github.com/lillyasviel/ControlNet-v1-1-nightly>, 2024. Accessed: DATE 2024-01-02. 3, 6
- [55] Lvmin Zhang, Chengze Li, Edgar Simo-Serra, Yi Ji, Tien-Tsin Wong, and Chunping Liu. User-guided line art flat filling with split filling mechanism. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9889–9898, 2021. 2
- [56] Lvmin Zhang, Chengze Li, Tien-Tsin Wong, Yi Ji, and Chunping Liu. Two-stage sketch colorization. *ACM Trans. Graph.*, 37(6):261, 2018. 2, 3
- [57] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, pages 3836–3847, 2023. 2, 3, 6, 8
- [58] Yuxin Zhang, Nisha Huang, Fan Tang, Haibin Huang, Chongyang Ma, Weiming Dong, and Changsheng Xu. Inversion-based style transfer with diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10146–10156, 2023. 3
- [59] Changqing Zou, Haoran Mo, Chengying Gao, Ruofei Du, and Hongbo Fu. Language-based colorization of scene sketches. *ACM Trans. Graph.*, 38(6), 2019. 2