OutfitCartoon: Dressing Up Your Anime Character With Diffusion Models

Yingying Zhang,[1] Haohan Zhang,[2] Jianshi Wu,[3]
[1]30920231154368, School of Information
[2]30920231154366, School of Information
[3]23020231154235, School of Information

Abstract

Recent advancements in Diffusion models and AI painting technologies have opened new avenues for exploring the customization of image generation. Some research has been conducted on character customization, while most of these works only involves real people, but not for cartoon characters. We introduce OutfitCartoon, a method to control large-scale diffusion models for cartoon character outfit customization, allowing users to customize the attire of their favorite animated characters. This work can be applied in many fields, such as providing inspiration for designers, facilitating game developers, and applying it in virtual reality and other areas. We aim to achieve personalized costume changes for stylized characters in multiple phases, starting with manually masked images and textual descriptions, and then focusing on processing results automatically.

Introduction

AI outfit customization is currently a hot research field, with an increasing number of people attempting to control generated images through the neural network. However, most of this kind of dress-up methods currently involve real people, and there are relatively few studies on cartoon character dress-up. It should be noticed that the world of anime character customization has witnessed a surge in popularity, with enthusiasts interested in customizing their favorite animated personas.

This work provides potential value across various domains, including offering creative inspiration to designers, aiding game developers, and finding utility in virtual reality and related fields. Yet cartoon character customization faces two primary challenges: character information conservation and keeping the feature of user-provided clothing. To address these challenges, we follow the up-to-date deep learning and image generation technologies and aim to make costume changes for cartoon characters.

In recent years, Diffusion Models(e.g.Stable Diffusion(Rombach et al. 2022)) has emerged as a groundbreaking technology, propelling image generation to new heights. The forward and reverse diffusion processes have unlocked exciting possibilities for data generation and manipulation. Due to the limited ability to finely control generated images, some efforts have explored various approaches to regulate the network of diffusion models, such as T2I-Adapter(Mou et al. 2023) and ControlNet(Zhang, Rao, and Agrawala 2023).

As for conditional image generation task, the newest state-of-the-art methods are EluCD_DDPM(Ma et al. 2023), ADM-G series(Zheng et al. 2022) etc.

In this paper, we aim to provide a framework called OutfitCartoon, which is inspired by ControlNet, with the aim of achieving customized outfitting for stylized characters, particularly anime and cartoon personas. We plan to unfold this work into multiple phases, starting with manual masks and textual descriptions for attire modifications and subsequently automate attire application, which could eliminate the need for manual masking in some way. Our work will be released as an extension for AUTOMATIC1111, which is a famous webUI framework for Stable Diffusion.

Related Work

Diffusion Models

Recently, diffusion models(Ho, Jain, and Abbeel 2020) have achieved tremendous success in the field of image generation. Before it came up, the most common models for image generation are Generative Adversarial Network(GAN) and Variational Autoencoder(VAE). In 2020, Denoising Diffusion Probabilistic Model(DDPM) was introduced, and like other generative models, it aims to generate target data samples from simple distribution noise.

The diffusion model always comprises two processes: the forward process and the reverse process. The forward process, also known as the diffusion process, transforms the image X_0 into a Gaussian distribution X_T through T iterations by adding random Gaussian noise. The reverse process can be used to generate data samples by recovering X_0 from X_T through multiple denoising steps.

Some methods were used to improve Diffusion and enhance its performance. For example, (Song, Meng, and Ermon 2020) introduced the Diffusion Denoising Implicit Model(DDIM). Building on the efficiency improvements of DDPM, DDIM achieved the effect of 1000-step sampling in just 50 steps. DDIM not only enable efficient sampling but also pioneered a deterministic sampling method, remi-

niscent of GAN Inversion, for image editing and generation.

Following this, OpenAI released "Classifier Guidance" (also known as Guided Diffusion)(Ho and Salimans 2022), which introduced a crucial strategy of guiding diffusion models to generate images using classifier-based guidance. Coupled with various other enhancements, diffusion models successfully surpassed the giants in the generative field. (Rombach et al. 2022) proposed a method, which could apply diffusion models in the latent space with powerful pretrained autoencoders. They introduced cross-attention layers into the model architecture, which makes it more efficient. (Dhariwal and Nichol 2021) improves the existing diffusion model architecture, and proposes a scheme that can balance the diversity and fidelity of image generation.

In April 2022, OpenAI unveiled DALL-E 2, which leveraged diffusion models and massive data to exhibit unprecedented levels of understanding and creative capabilities.

The release of Stable Diffusion(Rombach et al. 2022) in 2022 further propelled image generation, making it more aligned with human needs.

Controlling Image Diffusion Models

With the advancement of Diffusion models, there is an increasing focus on methods for controllable modulation. Significant progress has been made in this regard. MakeAScene(Gafni et al. 2022) encodes segmentation masks into tokens to control image generation. GLIGEN(Li et al. 2023) introduces new parameters for the diffusion model's attention layer to improve image generation. Text Inversion(Gal et al. 2022) and DreamBooth(Ruiz et al. 2023) allow fine-tuning of image diffusion models using a small number of user-provided example images to personalize the content in generated images. SpaText(Wong and Trevathan 2001) maps segmentation masks to local token embeddings.

Furthermore, Prompt-based image editing techniques provide practical tools for manipulating images using prompts. Voynov et al(Voynov, Aberman, and Cohen-Or 2023) propose an optimization method to adapt the diffusion process to sketches. T2I-Adapter(Mou et al. 2023) provide extra guidance to pre-trained text-to-image (T2I) models while not affecting their original network topology and generation ability. ControlNet(Zhang, Rao, and Agrawala 2023) add spatial conditioning controls to large, pretrained text-to-image diffusion models, and makes experiments to show that it may facilitate wider applications to control image diffusion models. Composer(Huang et al. 2023) introduces a novel generative paradigm that enables flexible control over output images, including spatial layout and color palettes, while maintaining compositional quality and model creativity. LoRA(Hu et al. 2021) proposes Low-Rank Adaptation, which freezes the pre-trained model weights and injects trainable rank decomposition matrices into each layer of the Transformer architecture, greatly reducing the number of trainable parameters for downstream tasks, and could be migrated to diffusion models. Besides, concurrent works are also examining a wide variety of ways to control diffusion models.

Method

Preliminary

The typical diffusion model consists of two processes: one is the forward process, which adds a small amount of Gaussian noise to the sample in T steps, the other is the corresponding backward process, which includes learnable parameters to estimate and eliminate noise in order to recover the input image. In this paper, we use Stable Diffusion as our exemplary base model to demonstrate how we utilize our OutfitCartoon workflow to perform character outfit changes.

Stable Diffusion includes several core components, such as a text encoder (using the CLIP's ViT-L/14 text encoder in Stable Diffusion) to convert user input prompt text into text embeddings, an Image Auto Encoder-Decoder model to encode the image into latent vectors or reconstruct the image from latent vectors, and a UNET structure for iterative denoising and multi-round prediction guided by text, transforming random Gaussian noise into image latent vectors. The samplers are responsible for carrying out the denoising steps, and much work(Karras et al. 2022) focuses on evaluating their properties. As for in cartoon character, we found that using PLMS(Liu et al. 2022) and UniPC(Zhao et al. 2023) sampler is always good for the task.

Our OutfitCartoon is a network structure that allows for outfit changes for anime characters based on text prompts or input images. We will first introduce the overall structure of OutfitCartoon, and then discuss how to use this network structure to achieve outfit changes and related details.

OutfitCartoon

OutfitCartoon uses ControlNet, combined with semantic information such as posture, depth, and canny in the image, to fix everything except the character's outfits. This aims to preserve the original content in the image and prevent significant differences between the generated image from the diffusion model and the original image. For input images, cartoon character-related images often have complex backgrounds with many details, so we implement methods to segment the images first. With powerful vision models capable of handling new image tasks, such as MODNet(Ke et al. 2022) and SAM(Kirillov et al. 2023), we can perform image segmentation before sending it to image generation models to synthesize outfits for characters. For the anime style, enhancing segmentation mask detection is necessary, since the composition of the images is generally complex. In practice, we use pre-trained MODNet network, which could do trimap-free portrait matting task in real time, and we implement SAM for image segmentation, extracting the background and non-clothing parts of the cartoon character. The segementation results are followed by the diffusion processes to generate the remaining "foreground" parts.

The whole network architecture is shown in Figure 1. Given an input image z_0, the diffusion process adds noise to the image and produce a noisy image z_t, where t represents the timestep of noise, which could be handled by noise schedular in denoise process. Given text prompt c_t, conditions c_f by ControlNet(such as depth map c_d, openpose c_p, and segmentation map c_s), together with LoRA connected

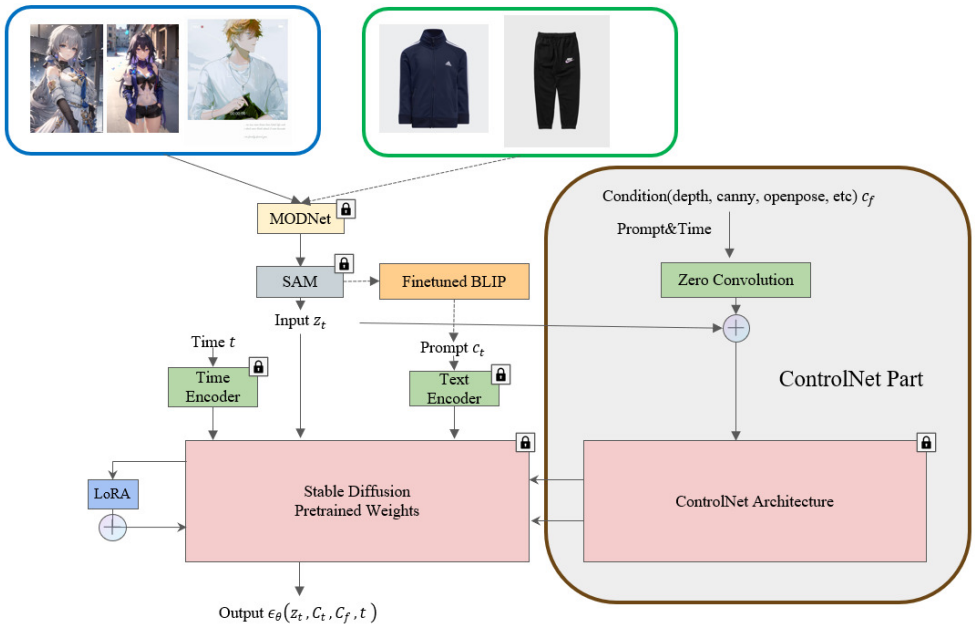

Figure 1: The network architecture of our OutfitCartoon. The LoRA and the BLIP parts are trainable and the other parts are fixed. The grey part shows how the ControlNet part is added to control the image generation process, and LoRA makes our network more familiar with the concepts of outfits. The conditions we give to ControlNet are depth map for hand generation, OpenPose input to keep posture, and segementation map to keep semantic information.

to the attention modules of UNet in diffusion models, the network learns a predictor ϵ_θ to predict the noise which is added to the original z_0. The loss function of the training process is

$$Loss = \mathbf{E}_{z_0,t,c_f,c_t,\epsilon\sim N(0,1)}\left[||\epsilon - \epsilon'_\theta(z_t, t, c_f, c_t)||_2^2\right], \quad (1)$$

where Loss is the learning objective for the OutfitCartoon, and ϵ'_θ means the denoising module conducted by LoRA and ControlNet.

Change Outfit Workflow

We make two workflows combined in an extension to do text-to-image and image-to-image tasks on cartoon characters' outfit generation and outfit changing in Automatic1111(A famous Stable Diffusion WebUI program), and make a small outfit dataset with around 100 images and target labels for LoRA to finetune stable diffusion. Since cartoon characters' outfit is always hard to extract and it may be difficult for the network to learn features, we use 60% real outfit images in our dataset and find that the LoRA has the ability to migrate the knowledge learning from real-world images to non-realistic style.

In text-to-Image workflow, we use text prompts and the original image with a cartoon character for image generation. OutfitCartoon can automatically extracting the foreground of the input image(in most time this is the character), then generate a segmentation map and a mask for inpainting, which could be edited by users. The image-to-image(Meng et al. 2021) method of the Diffusion model combined with Inpaint Anything(Yu et al. 2023) can be utilized to generate new clothing corresponding to a given text description using the original clothing mask.

As for the Image-to-Image workflow, we train a LoRA to guide diffusion models to generate clothes similar to the input image by adding bypass to the attention module. We still use pre-trained MODNet and SAM to segment the clothing in the input image, followed by using BLIP(Li et al. 2022) to extract features from the input image, to generate corresponding information and find similar prompts for what we trained for LoRA. We finetune the BLIP model with our outfit dataset, to make the output corresponded to the prompts we trained for LoRA. Training a model such as LoRA or lyCORIS(Yeh et al. 2023) to fine-tune the Diffusion Model can significantly improve the quality of generated outputs, since it helps LoRA guiding the Diffusion Models to memorize the concept of specific clothing. Additionally, it's important to address the issue of language drift. Language drift has been an observed problem in language models(Lu et al. 2020), and since Stable Diffusion is trained on a very large dataset, fine-tuning with a very small dataset can easily lead to knowledge forgetting. Specifying the fine-tuning range within the training set's text, i.e., CARTOONOUTFIT, can mitigate this type of forgetting.

Experiments

Qualitative Evaluation

Through Figure 1, we demonstrate the innovative clothing designer framework proposed in our research. This framework provides users with a new interactive platform, allow-

Figure 2: Our work can generate high-quality clothing designs based on given character images and text. The original image provided by the user is on the left, and custom prompts are placed above. The middle section demonstrates that after receiving user input, our framework provides personalized dressing functions for the characters in the original image without altering their posture or gestures.

ing them to easily upload original images and input custom prompts. Our framework is capable of providing personalized clothing change functions for the characters in the original images based on user input, while maintaining the characters' original poses and gestures. It is worth noting that different styles will produce distinctly different clothing effects and our work can perfectly fit different character images. It is evident from the figure that our framework has performed remarkably well in terms of controllability and fidelity of generation. This framework not only enhances user experience but also holds promise for offering some assistance to the clothing industry, providing new design and interaction methods for clothing designers and users.

For instance, under the condition of a single character image with the text prompt "bohemian style, layered clothing, earthy colors, fringe accessories, ethnic prints, natural fabrics, artistic designs" (1st row, 7th-10th columns), the generated image accurately depicts the given character wearing Bohemian-style clothing. The character in the image is dressed in typical Bohemian-style attire, which perfectly aligns with the elements of layered clothing, earthy colors, fringe accessories, ethnic prints, natural fabrics, and artistic designs. This seamless integration with the described character showcases our framework's ability to understand and accurately present specific clothing styles.

From Figure 2, we could find out that our system can process character images of various sizes, whether they are half-length or full-length, and perform outfit changes. At the same time, users can also define whether to retain the original background or generate a new one. Our framework utilizes pre-trained MODNet network and SAM for image segmentation, separating the background and non-clothing parts of the cartoon characters, and then uses ControlNet combined with Stable Diffusion to generate the remaining "foreground" parts. For example, in the first case, we successfully separated the characters from the background and generated new outfits while retaining the original background. In the second case, we demonstrate the system's ability to generate imaginative backgrounds based on user prompts. For instance, in the case of generating winter outfits (3rd row, 5th-7th columns), our system not only dressed the characters in winter clothing but also effectively simulated a snowy winter background.

Ablation 1: ControlNet vs Without ControlNet

The upper section provides a detailed introduction to our ControlNet. The images in the second columns display the dressing results using the ControlNet extension, while the images in the third and fourth columns show the dressing results without using ControlNet. It is clearly evident that the use of ControlNet preserves the pose of the character, resulting in better consistency with the original image. Conversely, not using ControlNet leads to catastrophic inaccuracies, resulting in significant differences from the original image. Figure 3 demonstrates that ControlNet performs better in capturing the posture of the clothing, accurately capturing both major body postures and minor posture details.

Ablation 2:Segmentation vs Without Segmentation

Thanks to the great contribution of Segment Anything, we

Figure 3: Qualitative results of ablation studies. The first column: input original image and prompt. The second column: results of using OutfitCartoon normally. The third and fourth columns: the effects after removing ControlNet. Differences are highlighted in the orange box. We can see that ControlNet can effectively preserve the posture of the original image. The fifth and sixth columns: the effects after removing the Segment algorithm. It is evident that without the Segment algorithm, the characteristics of the characters in the input original image are not preserved, and it appears to be a different person.

have implemented the function of segmenting characters. By segmenting the input picture, we are able to create masks to extract the desired features, such as the faces in the character images. Using the Segment Anything algorithm, we can extract the mask of the face to ensure that the pre-trained model does not alter the character's features while generating new clothing. In Figure 3, we conducted an ablation experiment. The images in the second column contain the original image, while the images in the fifth and sixth columns show the results without segmentation. It is clear that although ControlNet ensures the overall consistency of the posture, the character in the right image is obviously different from the one in the input image, indicating that the character has changed along with the change in clothing. Therefore, it can be seen that using the Segment Anything algorithm for mask image consistency control can effectively ensure consistency with the original image.

Summary and Future Work

In this paper, we propose a framework called OutfitCartoon, which utilizes large-scale diffusion models to achieve customized clothing for cartoon characters. Our framework aims to enable personalized clothing changes for stylized characters. Inspired by ControlNet, our framework aims to provide personalized dressing functions for cartoon characters based on user input, while maintaining the characters' original poses and gestures. Our network architecture includes the trainable LoRA and BLIP parts of ControlNet, with the remaining parts fixed. The ControlNet part controls the image generation process to make it more controllable, while LoRA familiarizes our network with the concept of outfits.

In future work, we plan to integrate the capability to generate character outfit changes based on specific clothing images provided by users. This additional feature will enhance the functionality of our framework, allowing users to visualize how characters would look wearing specific outfits.

Overall, our framework holds promise for providing new design and interaction methods for clothing designers and users, and offers potential applications in various domains such as virtual reality, gaming, and fashion design.

References

Dhariwal, P., and Nichol, A. 2021. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems* 34:8780–8794.

Gafni, O.; Polyak, A.; Ashual, O.; Sheynin, S.; Parikh, D.; and Taigman, Y. 2022. Make-a-scene: Scene-based text-to-image generation with human priors. In *European Conference on Computer Vision*, 89–106. Springer.

Gal, R.; Alaluf, Y.; Atzmon, Y.; Patashnik, O.; Bermano, A. H.; Chechik, G.; and Cohen-Or, D. 2022. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*.

Ho, J., and Salimans, T. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 6840–6851. Curran Associates, Inc.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Huang, L.; Chen, D.; Liu, Y.; Shen, Y.; Zhao, D.; and Zhou, J. 2023. Composer: Creative and controllable image synthesis with composable conditions. *arXiv preprint arXiv:2302.09778*.

Karras, T.; Aittala, M.; Aila, T.; and Laine, S. 2022. Elucidating the design space of diffusion-based generative models. *Advances in Neural Information Processing Systems* 35:26565–26577.

Ke, Z.; Sun, J.; Li, K.; Yan, Q.; and Lau, R. W. 2022. Modnet: Real-time trimap-free portrait matting via objective decomposition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 1140–1147.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. *arXiv preprint arXiv:2304.02643*.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, 12888–12900. PMLR.

Li, Y.; Liu, H.; Wu, Q.; Mu, F.; Yang, J.; Gao, J.; Li, C.; and Lee, Y. J. 2023. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22511–22521.

Liu, L.; Ren, Y.; Lin, Z.; and Zhao, Z. 2022. Pseudo numerical methods for diffusion models on manifolds. *arXiv preprint arXiv:2202.09778*.

Lu, Y.; Singhal, S.; Strub, F.; Courville, A.; and Pietquin, O. 2020. Countering language drift with seeded iterated learning. In *International Conference on Machine Learning*, 6437–6447. PMLR.

Ma, J.; Hu, T.; Wang, W.; and Sun, J. 2023. Elucidating the design space of classifier-guided diffusion generation. *arXiv preprint arXiv:2310.11311*.

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*.

Mou, C.; Wang, X.; Xie, L.; Zhang, J.; Qi, Z.; Shan, Y.; and Qie, X. 2023. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*.

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.

Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; and Aberman, K. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22500–22510.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Voynov, A.; Aberman, K.; and Cohen-Or, D. 2023. Sketch-guided text-to-image diffusion models. In *ACM SIGGRAPH 2023 Conference Proceedings*, 1–11.

Wong, M., and Trevathan, E. 2001. Infantile spasms. *Pediatric neurology* 24(2):89–98.

Yeh, S.-Y.; Hsieh, Y.-G.; Gao, Z.; Yang, B. B.; Oh, G.; and Gong, Y. 2023. Navigating text-to-image customization: From lycoris fine-tuning to model evaluation. *arXiv preprint arXiv:2309.14859*.

Yu, T.; Feng, R.; Feng, R.; Liu, J.; Jin, X.; Zeng, W.; and Chen, Z. 2023. Inpaint anything: Segment anything meets image inpainting. *arXiv preprint arXiv:2304.06790*.

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.

Zhao, W.; Bai, L.; Rao, Y.; Zhou, J.; and Lu, J. 2023. Unipc: A unified predictor-corrector framework for fast sampling of diffusion models. *arXiv preprint arXiv:2302.04867*.

Zheng, G.; Li, S.; Wang, H.; Yao, T.; Chen, Y.; Ding, S.; and Li, X. 2022. Entropy-driven sampling and training scheme for conditional diffusion generation. In *European Conference on Computer Vision*, 754–769. Springer.