ControlEdit: A MultiModal Local Clothing Image Editing Method

Di Cheng
202220001093@stu.bift.edu.cn
Yingjing Shi(✉)
20150015@bift.edu.cn
Shixin Sun
202220001094@stu.bift.edu.cn
Jiafu Zhang
202320001090@stu.bift.edu.cn
Weijing Wang
20230001092@stu.bift.edu.cn
Yu Liu
202320001091@stu.bift.edu.cn

School of Arts & Sciences,
Beijing Institute of Fashion
Technology, Beijing, China

Abstract

Multimodal clothing image editing refers to the precise adjustment and modification of clothing images using data such as textual descriptions and visual images as control conditions, which effectively improves the work efficiency of designers and reduces the threshold for user design. In this paper, we propose a new image editing method ControlEdit, which transfers clothing image editing to multimodal-guided local inpainting of clothing images. We address the difficulty of collecting real image datasets by leveraging the self-supervised learning approach. Based on this learning approach, we extend the channels of the feature extraction network to ensure consistent clothing image style before and after editing, and we design an inverse latent loss function to achieve soft control over the content of non-edited areas. In addition, we adopt Blended Latent Diffusion as the sampling method to make the editing boundaries transition naturally and enforce consistency of non-edited area content. Extensive experiments demonstrate that ControlEdit surpasses baseline algorithms in both qualitative and quantitative evaluations.The code and pretrained models will be available on GitHub. Check https://github.com/cd123-cd/ControlEdit

1 Introduction

Clothing image editing refers to the process of users making simple modifications to a given clothing image and obtaining a realistic modified physical image through algorithms. The clothing image editing model allows designers to interactively translate design concepts into real images. At the same time, ordinary users are allowed to communicate with professional

clothing designers to optimize the personalized clothing customization process using physical images as a reference. The majority of previous clothing image editing methods are based on GAN-based generative approaches [1, 8, 12, 16, 17, 28]. Recently, the FICE method[18] has combined GAN with the CLIP model to achieve semantic constraints, thereby achieving fine-grained content editing. Despite significant progress, training a new usable model is extremely dependent on dataset. Furthermore, the model's generation capability is restricted due to the scarcity of attribute text annotations. In addition, previous work has focused on attribute-guided image editing[1, 12, 16, 17, 18, 28] and sketch-guided image[8] editing. Although attribute-guided image editing can convey specific attributes of clothing such as style, color, and pattern, it may not provide enough geometric information to generate images consistent with what the user has in mind; sketch-guided image editing can assist in presenting the shape and layout of images, however, it is difficult to control semantic information such as image style. Hence, it is crucial to develop a multimodal clothing image editing method that combines sketches, text, and real images to enhance the clothing editing process.*f*

Recently, large-scale language image (LLI) models[7, 9, 14, 20, 21, 22, 25] have shown exceptional generation capabilities. These models allow for fine-tuning using various methods to adapt to downstream tasks in multiple domains[26]. There is relatively little work on multimodal clothing images editing in the fashion field, MGD[4] and RBI[15] have successfully fine-tuned pre-trained models to generate high-quality clothing images, making clothing image editing tasks possible. However, the commonly used LLI model introduces some randomness in the image generation process, which can result in slight differences in each generated image. The occurrence of "the slightest nudge causes the widest chain reaction" has a comprehensive impact on the final generated image. Therefore, it is particularly important to develop a controllable multimodal image editing method to optimize clothing design.

Figure 1: ControlEdit. Users can edit clothing by drawing conditional images.The first two rows of images are edited by regular users, while the last row is modified by professional fashion designers.

In this paper,we propose a local editing method based on sketches and text and define a

new task for multimodal conditioned fashion image editing. This method allows us to guide the generative process via multimodal prompts, maintaining the controllable, realistic, and plausible nature of the edited images(Fig.1). The key challenges of multimodal clothing image edit are: (1) It is difficult to collect sufficient paired real clothing images before and after modification for training. (2) The utilization of masks serves as the primary measure to maintain the integrity of non-edited regions, however, it introduces artifacts in the transition areas. Balancing the preservation of unchanged regions of clothing images alongside ensuring natural transitions between edited and non-edited regions while maintaining stylistic consistency poses a significant challenge. (3) Unlike image translation tasks, our task not only conducts domain translation from sketch to physical image, but also requires further reasonable fusion between the generated physical image and the source physical image.

The main contributions of this paper are as follows: (1) We propose ControlEdit, the first multimodal local editing method for clothing images, which is based on Controlnet. ControlEdit leverages sketches, natural language, and masked source images to guide image generation. Our editing process aligns with the typical practices of designers when making modifications.(2) We propose an inverse latent loss function, which optimizes the native Controlnet loss function and promotes consistency in non-edited area content. (3) We perform a mask fusion operation on the generated features and the source image features at each inference step in the latent space, avoiding issues such as unnatural pixel space mask transitions and inconsistent styles. (4) The above work has shown better image generation quality than the baseline model in benchmark testing.

2 Related works

GAN-based Clothe Image Edit In order to generate real clothing images, existing methods based on generative adversarial networks[6] usually map clothing control conditions to latent spaces and then perform clothing editing. Fashion++ [12]associates semantic segmentation maps with texture features and shape features. ADGAN[17] maps human attributes to latent space as independent code, and achieves attribute control through mixing and interpolation operations. FE-GAN[8] and FashionGAN[28] encode control images into the synthesized parsing map, which guide the generation of clothing image details. FashionTex[16] maps portrait, text and texture to latent space to obtain different latent vectors for manipulating image generation. FICE[18] utilizes pre-trained GAN [6]generators and CLIP models to implement semantic constraints. However, existing methods may encounter issues such as clothing image artifacts and lack of realism. This paper proposes an effective approach to address these issues by leveraging the robust generative capabilities of pre-trained models.

Diffusion-based Clothe Image Edit The rapid development of diffusion models[11] has been proven to surpass GANs, however, there is currently limited work on clothing image editing based on diffusion models. Text2Human[13] adds diverse text guidance to generate realistic texture portrait images based on human text for human body analysis. MGD[4] and [15] fine-tune pre-trained diffusion models to use reference images to complete missing areas while maintaining control condition guidance. DiffFashion[5] guides the denoising process through automatically generated semantic masks and pre-trained visual transformers (ViT)[2], allowing for appearance transfer while preserving structural information. Our approach differs from the aforementioned method, which emphasizes the use of textual descriptions and sketches as conditions for virtual try-on tasks. Instead, we focus on directly editing the garments themselves.

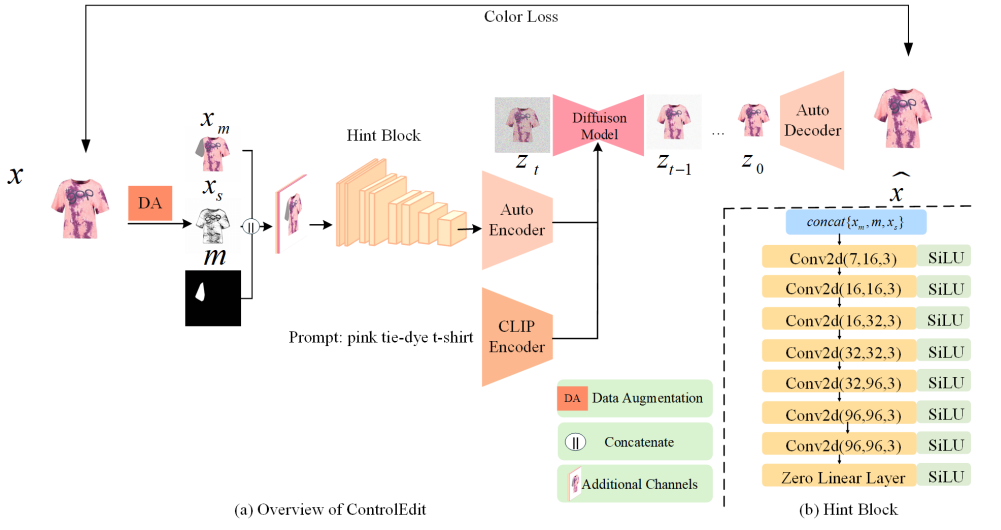

Figure 2: ControlEdit Network Architecture.

3 Method

During the process of local clothing image editing, given the image $x \in \mathbb{R}^{H\times W\times 3}$ to be edited, the user adds the mask $m \in \{0,1\}^{\mathrm{H}\times\mathrm{W}}$ on the image to simulate the modified area, where the value of 0 specifies the editable position, and the value of 1 ensures maximum consistent with x. The sketches drawn by users at masked positions are combined with the sketches extracted from the source image using Controlnet to obtain x_s through mask fusion operations, x_m representing the pre-editing state image. The training data is $\{(x_s, x_m, m, text), x\}$. Our aim is to conduct multimodal-based local image editing, wherein reference conditions are automatically integrated into the source image, ensuring that the resulting image appears controllable, realistic and plausible.

3.1 Preliminary

Controlnet Our ControlEdit is an extension of Controlnet[26], which is the fine-tuning method that copies the weights of LDM[21] to "trainable copy" and "locked copy". the locked copy retains the network capabilities learned from billions of images, while the trainable copy is trained on task specific datasets to learn conditional control, connected through zero convolution. Forward process: The extracted feature maps are fed into the autoencoder and converted into latent variable. Given the variance β, noise image z_0 is added until $z_T \sim N(0,1)$. The forward process is defined as follows:

$$q(z_t \mid z_{t-1}) = N\left(z_t; \sqrt{1-\beta_t} z_{t-1}, \beta_t I\right) \tag{1}$$

Reverse process: The reverse process can gradually remove noise by running reverse learning until a new sample is generated. Eq.2 allows for generating different reverse samples by changing the variance of the noise. Among them $\mu_\theta(z_t, t)$, $\Sigma_\theta(z_t, t)$ is the parameter for predicting Gaussian distribution.

$$p_\theta(z_{t-1} \mid z_t) = N(z_{t-1}; \mu_\theta(z_t, t), \Sigma_\theta(z_t, t)) \tag{2}$$

The loss function of Controlnet is shown in Eq.3, where text prompts c_t and c_f are conditional feature maps, and $\varepsilon_\theta(\cdot)$ is the denoising network.

$$L_{cldm} = E_{z_0,t,c_t,c_t,\varepsilon\sim N(0,1)}\left[\|\varepsilon - \varepsilon_\theta\left(z_t,t,c_t,c_f\right)\right)\|_2^2] \tag{3}$$

3.2 ControlEdit

The overall structure of ControlEdit is shown in Fig.2(a). Our task aims to generate target clothing images based on sketches, text, masks, and masked source images. We adopt Controlnet for initialization as the image prior to preserve the model's original controllability. Lacking a series of pre-edited real images, modified sketch images, and post-edited real images, we provide masked source images to the network to simulate pre-edited clothing images. The purpose is to allow the network to retain content in non-edited areas and provide color references for the generated areas when generating editing results; at the same time, in order to enhance the model's perception of editing positions, mask information is introduced to enable the network to better understand the spatial information of the target area of editing operations. Therefore, before the conditional features enter the denoising network, we extend the initial convolutional layer channel dimension of the conditional feature extraction network from 3 to 7 (i.e. 3+1+3), where $x_m \in R^{\{H,W,3\}}$ occupies three channels, $m \in \{0,1\}^{H\times W\times 1}$ occupies one channel, and $x_s \in R^{\{H,W;3\}}$ occupies three channels. Providing more parameter space for the network enables it to more fully express the complex relationship between input conditions and output images, improving the flexibility and expressive power of the model.

Figure 3: Masked Image Example.

Data augmentation The shape and size of clothing image editing are subject to randomness, so adopting conventional-shaped masks results in the models being limited to learning simple mappings. Inspired by the Paint by Example[23], we use the Bessel curve to sample 18 points and connect them to form a mask area of any shape, as shown in Fig.3. The generated mask area is closer to the actual editing operation, reducing the gap between training and testing, and enhancing the robustness.

Inverse latent loss function On the one hand, ControlNet method based on sketches encounters certain limitations in color restoration and detail preservation during model training, as it lacks the RGB information in the non-edited regions. On the other hand, the encoder of ControlNet performs multiple downsampling operations, which further aggravates the information loss. To ensure that the generated image aligns with the ground truth and that the network structure has RGB information of the source image, we introduce the masked source images into the feature extraction network. These images provide RGB information for non-edited regions, while masks prevent the leakage of the content that the model needs

to generate. The original Controlnet loss function is unable efficiently to bridge the gap between the editing and the non-editing domain. we propose the inverse latent loss function to force the editing model to pay more attention to maintaining the overall structure of the image and the consistency of the content in non-editing areas during the editing process. The image features predicted by the model are decoded into pixel space by the image decoder, and then we calculate L2 Euclidean distance between the decodeded image and the source image, which is as part of the total loss, as shown in Eq.5.We modify c_f in Eq.3 of L_{cldm} to be $c_f = E\left[\text{cat}\left(x_m, x_s, m\right)\right]$, where c_f represents the conditional feature map extracted by the encoder after concatenating the sketch, mask,and masked source image and $\hat{x}$ is the sampled image, which ensures the quality and realism of editing results.

$$L_{pix} = \|x - \hat{x}\|^2 \tag{4}$$

$$L = L_{cldm} + L_{pix} \tag{5}$$

Latent mask for sampling To further ensure the natural transition, we utilized the Blended

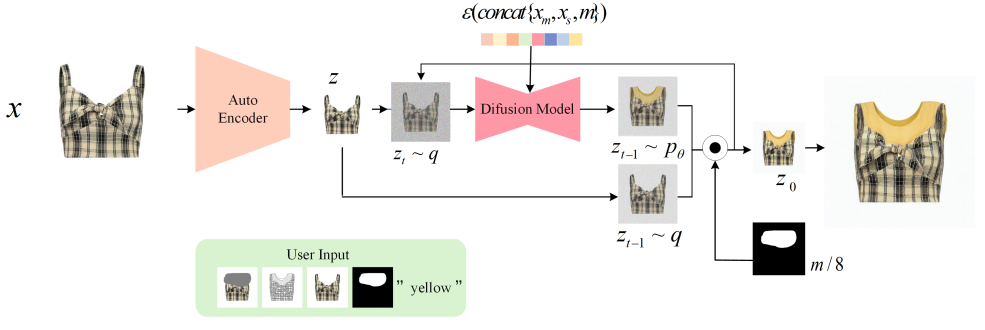

Figure 4: Image inference network structure.

Latent Diffusion[3] sampling method in the inference stage. Through modifying latent variables in each denoising step and forcing the parts outside the mask to remain unchanged, it ensures that the colors of non-editing areas naturally transition to the editing area, maintaining the global color consistency. As shown in Fig.4, in the denoising step, we adopt the features of *text* ,x_m, x_s, and m as conditioning inputs for the Unet to obtain the latent variables of the editing area. For reverse steps, z_{t-1}^{old} is sampled by the source image feature, and a certain level of noise is added to z according to Eq.6.

$$z_{t-1}^{\text{old}} \sim N\left(\sqrt{\bar{\alpha}_t} z_0, \left(1 - \bar{\alpha}_t\right) I\right) \tag{6}$$

Meanwhile, by using Eq.7 to process the noise of the original latent variable to the current noise level, we obtained the non-edited region latent variable with noise.

$$z_{t-1}^{\text{new}} \sim N\left(\mu_\theta\left(z_t, t\right), \Sigma_\theta\left(z_t, t\right)\right) \tag{7}$$

Subsequently, we mixed these two latent variables by adjusting the mask size according to Eq.8. It allows images to efficiently reduce obvious boundaries, making the overall effect more uniform.

$$z_{t-1} = m' \odot z_{t-1}^{\text{old}} + \left(1 - m'\right) \odot z_{t-1}^{\text{new}} \tag{8}$$

4 Experiments

4.1 Evaluations

Datasets We adopt the text prompts and real clothing images from the MGD[4] dataset, and extract clothing sketches from real clothing images by Controlnet. The MGD multimodal dataset consists of 11647 pairs of clothing images. We select 9394 images as training data and 2000 images as subsequent model evaluation data. In order to truly test the editing and robustness of the model, we make varying degrees of modifications to these 2000 sketches, such as adding patterns, deleting sleeves, and changing round necks into suit collars etc.

Quantitative evaluation In the task of image editing, the absence of precise evaluation metrics prompts the adoption of four metrics derived from the domain of generation for assessment purposes. FID[10] measures the distribution similarity between the real image and the generated image by comparing the mean and variance of image features. LPIPS[27] is used to evaluate the perceptual difference between two images. Pre_error[29] is the L2 distance between the non-edited regions of the generated image and the source image. The evaluation index for measuring the magnitude of image changes in non-edited areas. CLIP Score[19] is used to evaluate the semantic consistency between the generated image and the reference image.

4.2 Results and analysis

Baselines To our knowledge, this is the first time that diffusion models have been used for local image editing based on sketches and text. We choose three baseline models, and all methods are evaluated using the same 2000 images from the MGD dataset to sure fairness. 1) Controlnet. We use the edited sketch as a condition to represent the target content. 2)SD Inpainting. 3) Blended Latent Diffusion. 4)Uni-paint[24]. We use text prompts to represent the target content and masks to represent the generation area.

Figure 5: Qualitative comparison. SD Inpainting, Blended Latent Diffusion and Uni-paint synthesize clothing images driven by texts at the down side.

Table 1: Quantitative results of baseline model on 2000 images of size 512 x 512.

Method	FID↓	LPIPS↓	Pre_error↓	CLIP Score↑
Controlnet-Lineart[26]	10.689	0.1219	930.446	80.564
SD Inpainting [21]	9.970	0.0938	**1.322**	78.318
Blended Latent Diffusion [3]	7.49	0.1392	169.03	81.243
Uni-paint	8.669	0.0903	182.14	**81.994**
Ours	**4.569**	**0.0497**	80.672	81.684

Qualitative analysis We provide a qualitative comparison of these methods in Fig.5. Text-guided Blended Latent Diffusion, SD Inpainting and Uni-paint can generate images that match the description, however, it is difficult for regular language to specify fine-grained object appearances. Controlnet-Lineart focuses on translation between image domains with weak ability to maintain non-edited regions unchanged. Our method achieves unchanged non-editing areas, natural transitions, and generates areas that are loyal to the sketch and text conditions.

Quantitative analysis Tab.2 shows the quantitative comparison results. SD Inpainting achieves the lowest Pre_error score, indicating its ability to preserve information from conditional images, however, it focuses on inpainting task and lacks guidance on conditional information for image generation. Our method exhibits superior performance on most metrics and has significant advantages over existing methods. Especially, our method has the lowest scores in both FID and LPIPS, indicating better fidelity and perceptual similarity between the generated image and the real image. In addition, the Pre_error metric and CLIP-Score further confirm the effectiveness of our method in accurately reconstructing and preserving important features in generated images.

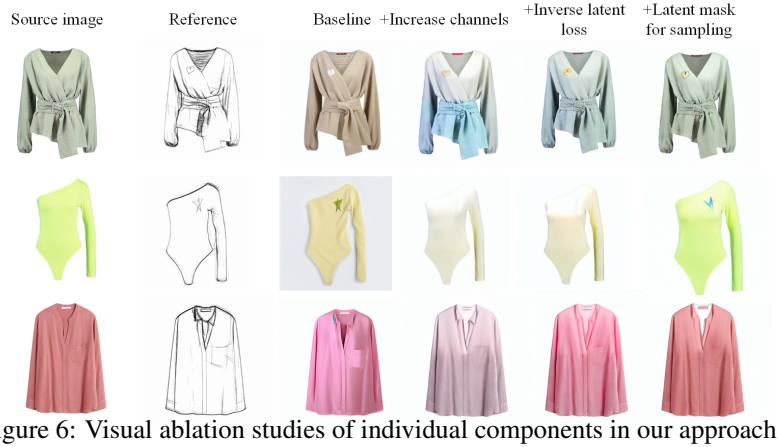

Figure 6: Visual ablation studies of individual components in our approach.

User study For models trained on the MGD dataset, participants are presented with source images, baseline-generated images, and images generated by our model. They are asked to select images based on three criteria: (1) realism, (2) color consistency, and (3) semantic consistency. We designed two approaches: 1) We select 10 participants to choose an image that best meets the aforementioned criteria. 2)We survey 100 participants, over

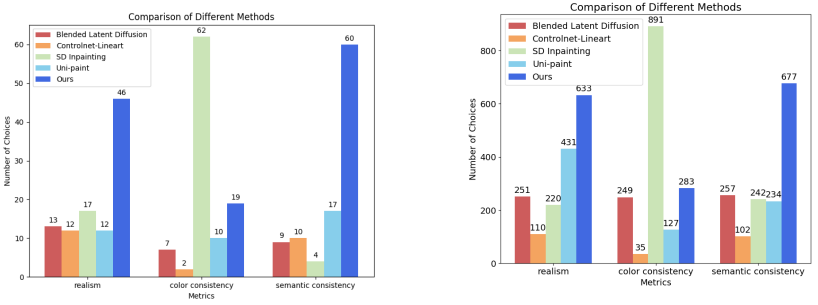

Figure 7: User study results. We compare our ControlEdit with three baselines.

Table 2: Quantitative results of ablation on 2000 images of size 512 x 512.

Method	FID↓	LPIPS↓	Pre_error↓	CLIP Score↑
Baseline	10.689	0.1219	930.446	80.564
+ Increase channels	5.773	0.0739	292.1	**81.704**
+ Inverse latent loss	5.388	0.0684	231.16	81.693
+ Latent mask for sampling	**4.569**	**0.0497**	**80.772**	81.684

half of whom are art majors, allowing them to select multiple images that meet the above criteria. As shown in Fig.7, whether single(left) or multiple(right) selections are allowed, our method exhibite superiority in human evaluation. Although the SD Inpainting method performe better in maintaining color consistency, it could not accurately generate semantic adjustments in clothing portions.

4.3 Ablations

To achieve multimodal local editing, we introduced three methods, namely Increase channels, Inverse latent loss, and Latent mask for sampling. 1) We represent Controlnet-Lineart as the baseline. 2) We have extended the Controlnet channels to express the complex relationship between input conditions and output images. 3) To ensure that the generated image matches the color of the source image, we add a new loss to the total loss. 4) To further ensure natural image transitions and color consistency, we use the Latent mask sampling method. We present the results in Tab.2 and Fig.6.The baseline solution generates images with a significant color difference between the generated image and the source image. By increasing channels, the image gradually approaches the source image in non-edited areas and transitions naturally, while the color style is inconsistent with the source image. When further adding loss, it alleviates the problem of non-editing area changes. Finally, using Latent mask further ensures that the color style of the generated image is consistent with that of the source image and the image transitions are natural, greatly improving the overall image quality and achieving the best performance.

5 Conclusion

We propose ControlEdit, which is the first pre-trained diffusion model used for local image editing methods based on sketches and text. Our method overcomes the problem of insufficient collection of clothing datasets through self-supervision. Our proposed loss function effectively preserves the details of non-edited regions. Numerous experiments have clearly demonstrated that ControlEdit outperforms existing methods in many metrics, achieving high-quality and realistic results. We hope that this work can serve as a solid baseline and contribute to supporting future research in the field of clothing image editing.

Acknowledgement

This work is supported in part by the Open Research Project of Hubei Provincial Engineering Research Center for Intelligent Textile and Apparel (2023HBITF01), the Strategic Cooperation Agreement between Cloudsky Information Technology Co. and Beijing Institute of Fashion Technology (Project No. H2024-98), the National Natural Science Foundation of China (Project No. 62062058), and the R&D Program of Beijing Municipal Education Commission (Project No. KM202210012002). I would like to thank Shuyang Gu, Shujie Liu, Qinwen Wei and Fang Yan for their valuable assistance with my research experiments and paper writing.

References

[1] Kenan E Ak, Joo Hwee Lim, Jo Yew Tham, and Ashraf A Kassim. Attribute manipulation generative adversarial networks for fashion images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10541–10550, 2019.

[2] Shir Amir, Yossi Gandelsman, Shai Bagon, and Tali Dekel. Deep vit features as dense visual descriptors. *arXiv preprint arXiv:2112.05814*, 2(3):4, 2021.

[3] Omri Avrahami, Ohad Fried, and Dani Lischinski. Blended latent diffusion. *ACM Transactions on Graphics (TOG)*, 42(4):1–11, 2023.

[4] Alberto Baldrati, Davide Morelli, Giuseppe Cartella, Marcella Cornia, Marco Bertini, and Rita Cucchiara. Multimodal garment designer: Human-centric latent diffusion models for fashion image editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23393–23402, 2023.

[5] Shidong Cao, Wenhao Chai, Shengyu Hao, Yanting Zhang, Hangyue Chen, and Gaoang Wang. Difffashion: Reference-based fashion design with structure-aware transfer by diffusion models. *IEEE Transactions on Multimedia*, 2023.

[6] Antonia Creswell, Tom White, Vincent Dumoulin, Kai Arulkumaran, Biswa Sengupta, and Anil A Bharath. Generative adversarial networks: An overview. *IEEE signal processing magazine*, 35(1):53–65, 2018.

[7] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. Cogview: Mastering text-to-image

generation via transformers. *Advances in Neural Information Processing Systems*, 34: 19822–19835, 2021.

[8] Haoye Dong, Xiaodan Liang, Yixuan Zhang, Xujie Zhang, Xiaohui Shen, Zhenyu Xie, Bowen Wu, and Jian Yin. Fashion editing with adversarial parsing learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8120–8128, 2020.

[9] Zhida Feng, Zhenyu Zhang, Xintong Yu, Yewei Fang, Lanxin Li, Xuyi Chen, Yuxiang Lu, Jiaxiang Liu, Weichong Yin, Shikun Feng, et al. Ernie-vilg 2.0: Improving text-to-image diffusion model with knowledge-enhanced mixture-of-denoising-experts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10135–10145, 2023.

[10] 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. *Advances in neural information processing systems*, 30, 2017.

[11] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[12] Wei-Lin Hsiao, Isay Katsman, Chao-Yuan Wu, Devi Parikh, and Kristen Grauman. Fashion++: Minimal edits for outfit improvement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5047–5056, 2019.

[13] Yuming Jiang, Shuai Yang, Haonan Qiu, Wayne Wu, Chen Change Loy, and Ziwei Liu. Text2human: Text-driven controllable human image generation. *ACM Transactions on Graphics (TOG)*, 41(4):1–11, 2022.

[14] Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, and Taesung Park. Scaling up gans for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10124–10134, 2023.

[15] 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.

[16] Anran Lin, Nanxuan Zhao, Shuliang Ning, Yuda Qiu, Baoyuan Wang, and Xiaoguang Han. Fashiontex: Controllable virtual try-on with text and texture. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–9, 2023.

[17] Yifang Men, Yiming Mao, Yuning Jiang, Wei-Ying Ma, and Zhouhui Lian. Controllable person image synthesis with attribute-decomposed gan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5084–5093, 2020.

[18] Martin Pernuš, Clinton Fookes, Vitomir Štruc, and Simon Dobrišek. Fice: Text-conditioned fashion image editing with guided gan inversion. *arXiv preprint arXiv:2301.02110*, 2023.

[19] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[20] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

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

[22] Chenfei Wu, Jian Liang, Lei Ji, Fan Yang, Yuejian Fang, Daxin Jiang, and Nan Duan. Nüwa: Visual synthesis pre-training for neural visual world creation. In *European conference on computer vision*, pages 720–736. Springer, 2022.

[23] Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. Paint by example: Exemplar-based image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18381–18391, 2023.

[24] Shiyuan Yang, Xiaodong Chen, and Jing Liao. Uni-paint: A unified framework for multimodal image inpainting with pretrained diffusion model. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, page 3190–3199, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701085. doi: 10.1145/3581783.3612200.

[25] Han Zhang, Weichong Yin, Yewei Fang, Lanxin Li, Boqiang Duan, Zhihua Wu, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. Ernie-vilg: Unified generative pre-training for bidirectional vision-language generation. *arXiv preprint arXiv:2112.15283*, 2021.

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

[27] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

[28] Shizhan Zhu, Raquel Urtasun, Sanja Fidler, Dahua Lin, and Chen Change Loy. Be your own prada: Fashion synthesis with structural coherence. In *Proceedings of the IEEE international conference on computer vision*, pages 1680–1688, 2017.

[29] Zixin Zhu, Xuelu Feng, Dongdong Chen, Jianmin Bao, Le Wang, Yinpeng Chen, Lu Yuan, and Gang Hua. Designing a better asymmetric vqgan for stablediffusion, 2023.