Realistic Human Generation with Controllable Poses Using 3D Priors

Ruifeng Bai[†]**, Xiaohang Liu**[†]**, Haozhe Jia, Wei Zhang, Changpeng Yang, Di Xu**

Huawei Cloud
bairuifeng4@huawei.com, liuxiaohang3@huawei.com

Abstract

In the past decade, generative artificial intelligence has made significant progress, especially in image generation. Realistic photos can be created in seconds with straightforward text prompts. However, when it comes to the human body generation, precise controllability of desired poses is still an open problem. This could be even more tricky when multi-view consistency is considered. To this end, we propose to tackle such 2D problems with the help of 3D priors. Specifically, Stable Diffusion is applied to generate high-quality images, while the Low-Rank Adaptation (LoRA) takes charge of the character style. As an essential connection between 2D and 3D, we introduce SMPL as a novel prior to explicitly control the generated poses. The experimental results show that the proposed method overperforms the state-of-the-arts, effectively generating realistic human photos with sophisticated poses.

Introduction

With the rapid development of AI image generation, all kinds of images can be generated from text prompts or visual cues. Among them, generating human portraits is one of the most import but challenging topic. The goal is to generate digital human with real-life appearance, expressions and poses. In the early stage of human image generation, the Generative Adversarial Network (GAN) (Goodfellow et al. 2014) is the mainstream of the generation model (Karras et al. 2020; Liao et al. 2022), and the generator structure is used to effectively generate images, but it is easy to cause model collapse during the training process. Most recently, the diffusion model has become more and more popular as a successful generation method (Saharia et al. 2022; Rombach et al. 2022; Ramesh et al. 2022; Nichol et al. 2021). This method uses the Markov chain (Dhariwal and Nichol 2021) learning mechanism to model the image through the Gaussian distribution (Do 2008) to achieve image denoising. Similarly, the breakthrough in image generation models is mainly driven by text-guided diffusion models. These diffusion-based text-to-image generation models (Yu et al. 2022; Chang et al. 2023; Ramesh et al. 2022) have made impressive improvements in image quality and have outperformed GAN (Kang et al. 2023; Dhariwal and Nichol 2021) in performance.

In particular, for human pose control, GAN is usually used, but when the person undergoes large pose changes, it is impossible to capture the reasonable texture mapping between the original image and the target image (Zhang et al. 2022). Therefore, this paper adopts Stable Diffusion (Rombach et al. 2022) as the basic model, and uses LoRA (Hu et al. 2021) to control the character style, and trains ControNet (Zhang, Rao, and Agrawala 2023) with the human image datasets. The Stable Diffusion model can effectively understand and extract text semantics to generate images of related concepts. In text-to-image generation, ControlNet makes the generated image of the diffusion model more controllable. The trained ControlNet can further adjust the diffusion model according to the image information such as canny edge map, depth map and openpose (Cao et al. 2017), and control the human body posture, edge feature, front and back position relationship of the generated image. However, when controlling the posture of the human body only through openpose, the capture of human details and posture is not sufficient. Therefore, we introduce the SMPL (Loper et al. 2023) as a 3D prior to explicitly control the human pose, thereby more effectively obtaining the human pose and appearance representation.

Preliminaries

Stable Diffusion. Stable Diffusion (Rombach et al. 2022) is a text-controlled image generation model based on diffusion structure, which uses latent space to significantly reduce the computational resources required for high-resolution training and reasoning. Stable Diffusion is mainly composed of three parts: 1) Variational autoencoder (VAE) (Kingma and Welling 2013), which includes an encoder and a decoder. The encoder effectively preserves the deep image features while converting the image into a low-dimensional latent space representation of UNet, and then the decoder creates the image based on the representation of the latent space. 2) UNet is an encoder and decoder based on residual module. The encoder realizes image compression, and the decoder decodes low-resolution images into high-resolution images. 3) Text Encoder encodes the input text as token embeddings, converts the input text into the meaning that UNet

[†]These authors contributed equally.

can comprehend, and then generates an image that conforms to the text description. The Stable Diffusion model has also been extended to text-based image operations and supports local and global editing as well as personalized operations. In order to facilitate model loading and image generation, this paper uses Stable Diffusion as the basis of the current image generation framework (Figure 1).

LoRA. LoRA (Hu et al. 2021) is mainly used to solve the problem of fine-tuning large models. Previously, the adjustment of Stable Diffusion is slow and challenging. Considering that the calculation of the gradient does not require the model weight, the LoRA introduces a trainable layer in each Transformer block, which greatly reduces the number of training parameters. The LoRA fine-tuning is more efficient and less computationally intensive, while maintaining the same quality level as the full model fine-tuning.

ControlNet. ControlNet (Zhang, Rao, and Agrawala 2023) is largely dependent on UNet and belongs to the model of replication-diffusion UNet. The model uses convolutional layers to connect new conditional inputs and the output of each layer as an encoder to control the image structure generated. The ControlNet model controls the picture by adding more conditions to the Stable Diffusion model, so as to accurately adjust the final generated content, and it is easier to adjust the strong randomness generation result of the diffusion model.

SMPL. SMPL (Loper et al. 2023) is a parameterized human body model, which is a method of human body modeling by learning a large number of different human body databases. It can realize the modeling of human bodies with different postures by changing the input of parameters. The model has two input parameters: body shape parameter β and pose parameter θ. The body shape parameter β is a 10-dimensional vector, which describes the body shape characteristics of the human model. The post parameter θ is a 75-dimensional vector, which is used to describe the changes of human joints. By specifying the body shape parameter β and the pose parameter θ, and acting on the average template to deform the body shape and movement, the specified human body model is reconstructed. The formula is expressed as:

$$M(\overrightarrow{\beta}, \overrightarrow{\theta}) = W(T_P(\overrightarrow{\beta}, \overrightarrow{\theta}), J(\overrightarrow{\beta}), \overrightarrow{\theta}, w) \tag{1}$$

$$T_P(\overrightarrow{\beta}, \overrightarrow{\theta}) = \overline{T} + B_S(\overrightarrow{\beta}) + B_P(\overrightarrow{\theta}) \tag{2}$$

In the formula: W is a mixed skin linear equation. J is a function that corresponds the body shape parameter β to the bone joint point. B_P is the function of mapping the attitude parameter θ to the corresponding point of the model. B_S is a function of β mapping to the corresponding point of the model. w is the mixed weight of each joint.

In this study, Human Bodies in the Wild (HBW) (Choutas et al. 2022) dataset and 3D Poses in the Wild (3DPW) (Von Marcard et al. 2018) dataset are used for training dataset. Only the full body image data of participants in the laboratory white background are extracted for HBW dataset. At the same time, only single video sequences in different scenes are selected from 3DPW. As shown in Figure 1, we train the LoRA on the HBW dataset, including the training of the peoples and the training of the faces. Through the trained LoRA, the optimization of the Stable Diffusion large model is realized, and then the ControlNet model conditions are applied to adjust the shape of the peoples. Among them, the ControlNet model is trained by obtaining pose control images corresponding to 3D Prior (SMPL) from the 3DPW dataset.

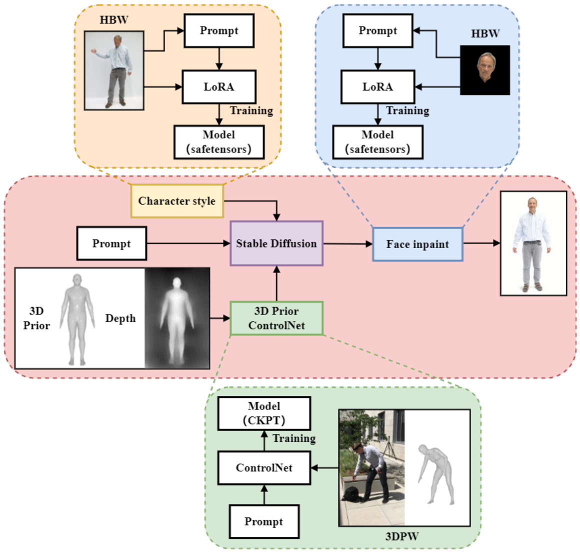

Figure 1: The overall framework. Character style: The background-free HBW data completes the training of LoRA and realizes the adjustment of character style. Face inpaint: The LoRA trained by the background-free HBW face data accurately regulates the human face. 3D Prior ControlNet: The pose image obtained by 3D prior (SMPL) completes the training of ControlNet and realizes the control of human pose. 3D Prior: The pose control image is obtained from 3D prior (SMPL). Depth: The depth image obtained from the pose control image.

Experiments

LoRA Training. We select different participants from the HBW dataset to train LoRA. For each set of data, u2net_human_seg (Qin et al. 2020) is first used to remove the background of the person, preventing LoRA from learning irrelevant background information. Then, the BLIP (Li et al. 2022) algorithm is used to extract prompt for each RGB photo. In order to make LoRA learn more detailed clothing and perspective features, the description words of clothing and perspective are manually refined on the basis of the prompt words generated by BLIP. The base model of LoRA training is realisticVisionV51_v51VAE, which is trained for 20 epochs.

In the field of human face generation, the face is an essential factor in determining the quality of the generated picture. Therefore, to improve the realism and accuracy of generated human faces, we have adopted a series of techniques to recompose the faces after generation. Specifically, the YOLOv8 (Lou et al. 2023) algorithm is used for face detection, and the face LoRA is trained using the domesticated face data. The Faceparsing (Lee, Bhattarai, and Kim 2021) algorithm is also used to segment the face in the iamge, and the face data is separated from the image, so as to process the face more accurately. In addition, in order to maintain the consistency of the face and the full body, the prompt of the face is adjusted accordingly. With these techniques, we have successfully improved the quality of the face in generated human images, leading to further improvement in the overall generated effect.

ControlNet Training. In the training process of ControlNet, in order to ensure the accuracy and stability of model training, the u2net_human_seg segmentation algorithm is also used to separate the background and human, which effectively eliminates the interference of the background for model training. Then, the BLIP algorithm is used to extract prompt. Importantly, the corresponding 3D prior is obtained by removing the background image, and then the pose control image obtained by the 3D prior is used as a new control condition. Finally, the obtained prompt, and the removal of background images and pose control images are used for fine-tuning of ControlNet. In this way, the training effect of ControlNet is effectively improved, so as to achieve more accurate and reliable human pose estimation.

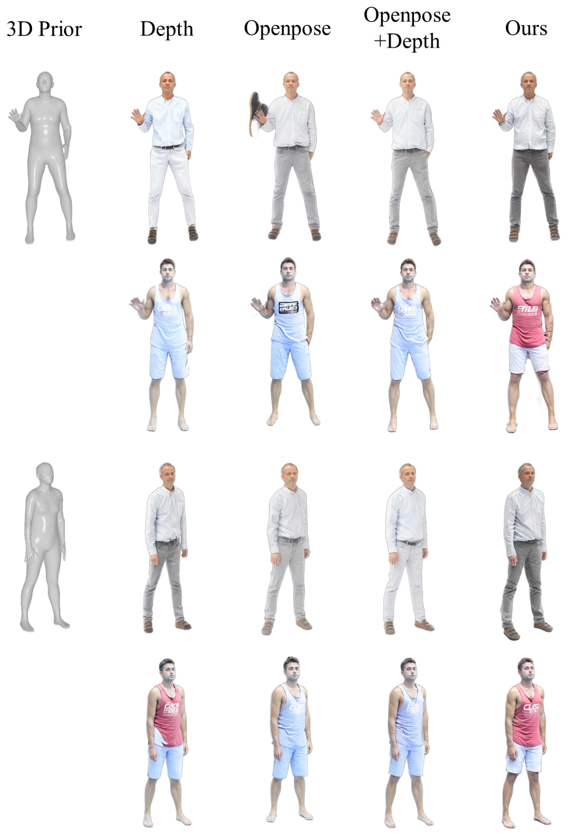

Figure 2: Qualitative comparison of text-to-image generation.

In this paper, the trained LoRA is used to optimize the Stable Diffusion large model, and the ControlNet model is used to adjust the human body pose, and then the human images of different postures are generated. Further, this study conducted experiments on the text-to-image and image-to-image. Through the comparison of the results generated by different generation methods, the superiority of this study is further explored. It is worth mentioning that in the experiment of image-to-image, not only the generation of single view is considered, but also the generation of multi-view is considered, which provides more comprehensive experimental results.

Results

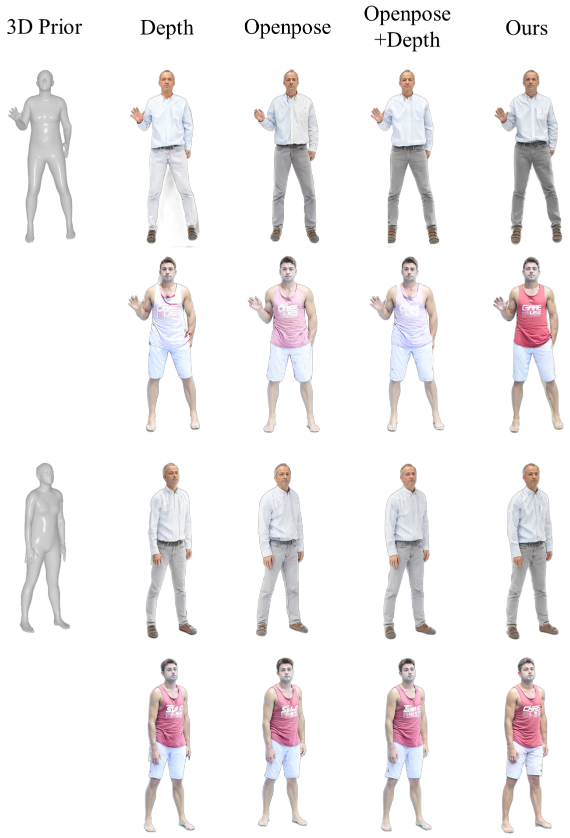

Figure 3: Qualitative comparison of image-to-image generation.

Text-to-Image. Firstly, the data of the 4th and 13th participants in HBW are selected for comparative experiments. By adjusting the parameters corresponding to the 3D prior, a new pose is obtained as the control condition of human pose. Then, the corresponding depth image and openpose iamge are obtained from the pose control image. Finally, under the same prompt, our results are qualitatively compared with depth, openpose and openpose+depth, respectively. The qualitative results are shown in Figure 2. Compared with the above comparison methods, our method has a significant improvement in human pose and appearance. Specifically, it surpasses depth, openpose and openpose+depth in dress, and maintains consistency with the original image. At the same time, it is also superior to other comparison methods in hand generation details. In particular, in terms of foot posture, our method is superior to openpose, 3D Prior ControlNet effectively controls the pose of the human and generates high-quality human images.

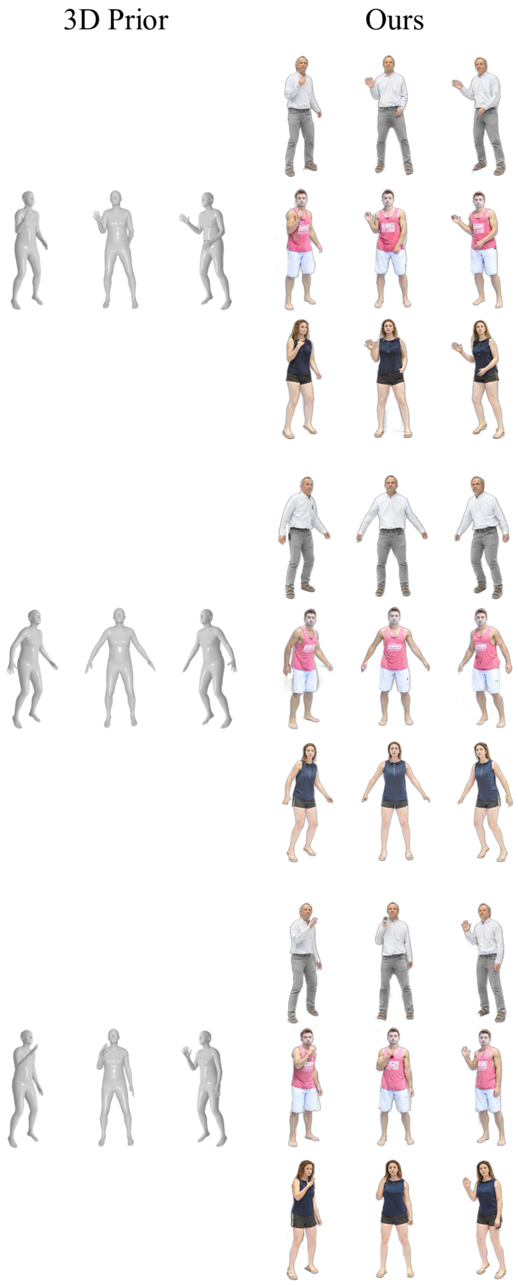

Figure 4: Multi-view images generation.

Image-to-Image. In the experiment of image-to-image, we conduct qualitative and quantitative analysis. Similarly, under the same prompt, the 4th and 13th participants are selected for the experiment. See Figure 3, through qualitative comparison, it is found that the results of our method are optimal. Specifically, it significantly surpasses depth in dress, and is significantly better than openpose in foot and hand pose control, with more detailed features. Quantitative results are presented in Table 1, Frechet Inception Distance (FID), Root Mean Squared Error (RMSE), Peak Signal Noise Ratio (PSNR) and Structural Similarity Index (SSIM) are used for comparison. Our method achieves the best RMSE and PSNR values, and the dress and human image are closer to the original image. On other metrics, our method is comparable to the advanced opnepose and competitive in posture adjustment accuracy and consistency.

	FID ↓	RMSE↓	PSNR↑	SSIM↑
Depth	150.64	47.22	14.85	0.94
Openpose	**123.95**	42.66	15.74	**0.96**
Openpose+Depth	127.51	44.60	15.39	0.95
Ours	125.77	**41.16**	**16.04**	**0.96**

Table 1: Quantitative comparison.

For the generation of multi-view human images, it is necessary to obtain pose control images at different angles ($-45°$, $0°$, $45°$) by adjusting different camera parameters. In addition, the 35th participant data in HBW is also added for the experiment. As shown in Figure 4, our method can effectively maintain the consistency of pose on different views, and achieve a unified effect on clothing.

Conclusion

In this paper, SMPL is introduced as a 3D prior to explicitly control the human body poses, increasing the stability and fidelity of human generation. We use Stable Diffusion as the basic framework for human image generation, and train LoRA to control character style, then use 3D prior to train ControlNet for desired human poses. As shown in the experiments, our method is able to maintain the consistency in both the appearance and pose of particular characters. To sum up, we propose a novel method for the generation of human images, which has wide applications in media, e-commerce and video production.

References

Cao, Z.; Simon, T.; Wei, S.-E.; and Sheikh, Y. 2017. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7291–7299.

Chang, H.; Zhang, H.; Barber, J.; Maschinot, A.; Lezama, J.; Jiang, L.; Yang, M.-H.; Murphy, K.; Freeman, W. T.; Rubinstein, M.; et al. 2023. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*.

Choutas, V.; Müller, L.; Huang, C.-H. P.; Tang, S.; Tzionas, D.; and Black, M. J. 2022. Accurate 3D body shape regression using metric and semantic attributes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2718–2728.

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

Do, C. B. 2008. The multivariate Gaussian distribution. *Section Notes, Lecture on Machine Learning, CS*, 229.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. *Advances in neural information processing systems*, 27.

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*.

Kang, M.; Zhu, J.-Y.; Zhang, R.; Park, J.; Shechtman, E.; Paris, S.; and Park, T. 2023. Scaling up gans for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10124–10134.

Karras, T.; Laine, S.; Aittala, M.; Hellsten, J.; Lehtinen, J.; and Aila, T. 2020. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8110–8119.

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Lee, J.; Bhattarai, B.; and Kim, T.-K. 2021. Face parsing from RGB and depth using cross-domain mutual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1501–1510.

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.

Liao, W.; Hu, K.; Yang, M. Y.; and Rosenhahn, B. 2022. Text to image generation with semantic-spatial aware gan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 18187–18196.

Loper, M.; Mahmood, N.; Romero, J.; Pons-Moll, G.; and Black, M. J. 2023. SMPL: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, 851–866.

Lou, H.; Duan, X.; Guo, J.; Liu, H.; Gu, J.; Bi, L.; and Chen, H. 2023. DC-YOLOv8: Small-Size Object Detection Algorithm Based on Camera Sensor. *Electronics*, 12(10): 2323.

Nichol, A.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; McGrew, B.; Sutskever, I.; and Chen, M. 2021. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*.

Qin, X.; Zhang, Z.; Huang, C.; Dehghan, M.; Zaiane, O. R.; and Jagersand, M. 2020. U2-Net: Going deeper with nested U-structure for salient object detection. *Pattern recognition*, 106: 107404.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2): 3.

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.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E. L.; Ghasemipour, K.; Gontijo Lopes, R.; Karagol Ayan, B.; Salimans, T.; et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35: 36479–36494.

Von Marcard, T.; Henschel, R.; Black, M. J.; Rosenhahn, B.; and Pons-Moll, G. 2018. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *Proceedings of the European conference on computer vision (ECCV)*, 601–617.

Yu, J.; Xu, Y.; Koh, J. Y.; Luong, T.; Baid, G.; Wang, Z.; Vasudevan, V.; Ku, A.; Yang, Y.; Ayan, B. K.; et al. 2022. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2(3): 5.

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

Zhang, P.; Yang, L.; Lai, J.-H.; and Xie, X. 2022. Exploring dual-task correlation for pose guided person image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7713–7722.