Improving the Accuracy of Pose-Guided Human Image Generation Via Multi-stage ControlNet Fine-Tuning

Jiu YI[†] Hiroshi WATANABE[†]

[†] Graduate School of Fundamental Science and Engineering, Waseda University

Abstract Pose-Guided human image generation aims to synthesize realistic human image conditioned on a target pose. Despite recent advances using ControlNet to integrate additional condition for structural control, the original method still struggles precise pose control. In the paper, we propose multi-stage fine-tuning method to improve the pose accuracy.

1. Introduction

Pose-Guided Human Image generation is a task which generates new human-centric images based on conditional inputs, such as text and pose image. Early methods relied on GANs [1] and VAEs [2], which the generated image suffer from poor quality and weak pose alignment. With the emergence of SD [3], the quality of generated image improved significantly. Nowadays, the popular method widely used is called ControlNet, which introduces an additional learnable branch to the frozen SD for conditional generation.

However, ControlNet [4] still struggles to achieve precise pose control on human image generation, meaning the generated image may not align well with the pose condition. Therefore, we propose a multi-stage fine-tuning method for ControlNet to improve pose accuracy. We conduct extensive evaluations of our method, which shows that it significantly improves pose fidelity while keeps the original generation quality compare with the baseline.

2. Related Work

2.1 ControlNet

ControlNet [4] is a neural network that allows diffusion model to integrate additional conditions for more precise structural control during the generation, such condition can be human pose, depth map, sketch and so on. It copied part of backbone of SD [3] and attaches it as an extra branch to the original model. During the training, the parameters of original SD is frozen, while only the extra branch is trained. This design preserves the generation capability of pre-trained SD, while enables the model to learn new conditions with relatively low computation cost. However, it still has problems of generating fine details of human body or handling complex pose.

2.2 Heatmap-guided denoising loss

Heatmap-guided denoising loss is originally invented from HumanSD [5], which is used to fine-tune SD model on pose condition. The target is to assign higher weights in pose-relevant regions during the training by creating a heatmap mask in the latent space. Therefore, the SD model can focus on the pose area instead of background during generation so that improves the pose alignment.

The loss is designed as follows:

$$\boldsymbol{L_h = E_{t,z,\varepsilon}\left[\left\|W_a \cdot \left(\varepsilon - \varepsilon_\theta\left(\sqrt{(\bar{\alpha}_t)}z_0 + \sqrt{(1-\bar{\alpha}_t)}\varepsilon, c, t\right)\right)\right\|^2\right]}, \quad \text{(i)}$$

where $W_a = w \cdot H_E + 1$, H_E is the heatmap mask.

Our method is inspired from this custom denoising loss and attempt to apply it into fine-tuning ControlNet to further improve pose accuracy of generated human image.

3. Proposed Method

We propose a multi-stage fine-tuning method for ControlNet to improve the pose alignment between the generated human image and pose input. As shown in Fig.1., We freeze the parameters of VAE encoder and SD [3] model, only update the parameters of ControlNet during the fine-tuning. We apply two different loss function in the separate stage. In the Stage 1, we adopt the original latent-space denoising loss from SD. The goal is to obtain a converged ControlNet that enables the generated image to roughly follow the input pose. This stage focuses on making the model responsive to diverse pose conditions. In the Stage 2, we apply heatmap guided denoising loss. The training objective is to refine the model to accurately follow the input pose. We apply a heatmap mask in the latent space to assign higher weights to pose-relevant regions. This encourages model to focus more on human structure.

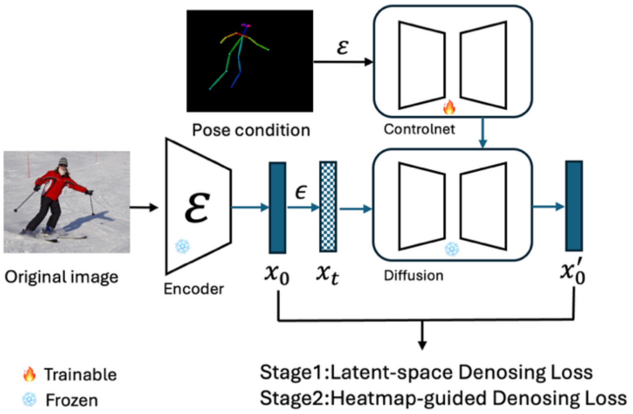

Fig. 1. Overview of multi-stage ControlNet Fine-tuning

4. Experiment

We conduct experiments on the Captioned COCO-Pose dataset [6] with 61.4k training and 2.69k evaluation triplets, using Stable Diffusion 1.5 as the base model. Training is done in two stages. Stage 1 uses the original denoising loss for 4 epochs. Stage 2 uses a heatmap-guided loss for 2 epochs. Both stages use batch size 1, learning rate 5×10^{-7}, and gradient accumulation step 1.

To evaluate the effectiveness of our proposed multi-stage fine-tuning method, we compare the models from both stages with the baseline model, ControlNet-OpenPose. The evaluation is based on three main criteria: OKS for pose accuracy, LPIPS for assessing image quality, and CLIP score for measuring text-image alignment. We conduct both qualitative and quantitative evaluations of our method. Table 1 presents the quantitative comparison with the baseline, showing that our Stage 2 model significantly improves pose accuracy while slightly enhancing the original model's generation capability.

Table 1: Quantitative comparison with the baseline in terms of pose accuracy, image quality, and text-image alignment

Model	CLIP Score ↑	LPIPS ↓	OKS ↑
Stage1	32.3476	0.7762	0.6853
Controlnet-Openpose	31.3786	0.7956	0.7186
Stage2	**31.9787**	**0.7657**	**0.7857**

For qualitative evaluation, we compare generated images from Stage 1, Stage 2, and the baseline using the same text prompt and pose. As shown in Fig. 2, the Stage 2 model achieves better pose alignment, with keypoints more closely matching the input. Fig. 3 further shows improved text-image alignment, only the Stage 2 model correctly includes the bicycle mentioned in the prompt.

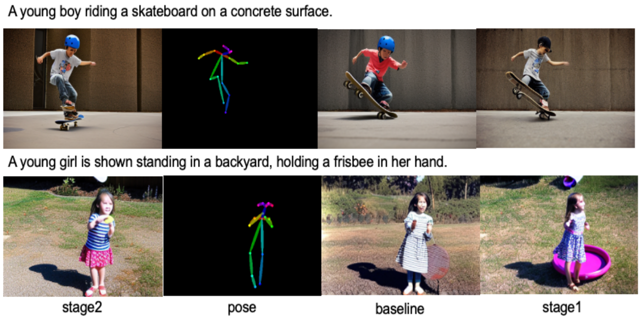
A young boy riding a skateboard on a concrete surface.
A young girl is shown standing in a backyard, holding a frisbee in her hand.
stage2
pose
baseline
stage1

Fig. 2. Qualitative results illustrating pose accuracy

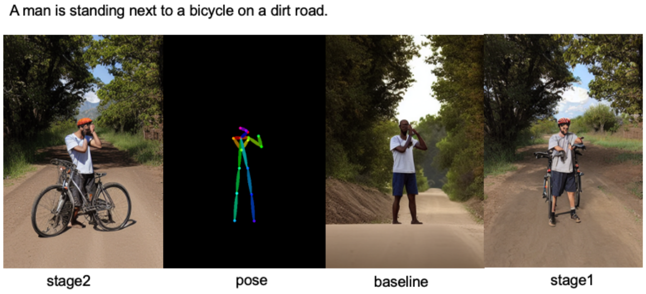

Fig. 3. Qualitative results illustrating text-image alignment

5. Conclusion

We propose a multi-stage fine-tuning method for ControlNet to improve pose accuracy in generated human images. Stage 1 uses the original latent denoising loss, while Stage 2 introduces a heatmap-guided loss for better pose alignment. Extensive evaluations show significant improvements in pose fidelity over the baseline.

Reference

[1] Goodfellow et al., "Generative adversarial networks," Commun. ACM, vol. 63, no. 11, pp. 139–144, 2020.

[2] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," arXiv:1312.6114, 2013.

[3] Rombach et al., "High-resolution image synthesis with latent diffusion models," CVPR, pp. 10684–10695, 2022.

[4] L. Zhang and M. Agrawala, "Adding conditional control to text-to-image diffusion models," arXiv:2302.05543, 2023.

[5] X. Ju et al., "HumanSD: A native skeleton-guided diffusion model for human image generation," ICCV, pp. 15988–15998, 2023.

[6] Liming CV, "Captioned COCOPose dataset," HuggingFace.Available:https://huggingface.co/datasets/limingcv/Captioned_COCOPose, 2023.

† 早稲田大学大学院 基幹理工学研究科 情報理工・情報通信専攻
〒162-0072 東京都新宿区大久保 3-14-9 早大 66-401
Phone: 03-5286-2509, E-mail: yijiu@fuji.waseda.jp