PostureCraft: Controllable Human Pose Generation with Diffusion Models

Penghui Qi*
National University of Singapore
penghuiq@comp.nus.edu.sg

Yiyao Huang*
National University of Singapore
e1322639@u.nus.edu

Qiang Wang*
National University of Singapore
e1327918@u.nus.edu

Weiqi Huang*
National University of Singapore
e1327928@u.nus.edu

Abstract

In recent years, diffusion models have emerged as a leading paradigm in generative artificial intelligence, producing high-quality images across a wide range of applications. However, the ability to precisely control specific attributes of generated images, such as the posture of human figures, remains a significant challenge. This paper introduces PostureCraft, a novel approach that integrates segmented depth map and depth annotated OpenPose control of human posture into the denoising process of diffusion models. By utilizing the expressive depth map and filtering out irrelevant information by segmentation, our fine-tuned model enables the generation of images with precise control over the posture of individuals. Additionally, our proposed depth annotated OpenPose control enhances the fidelity of hand representations, which is a well-known issue in human image generation. Our methodology extends the capabilities of diffusion models beyond text-based conditions, allowing for the generation of images that adhere to specific posture and gesture requirements. Models are available at `https://huggingface.co/akina/5340_project`.

1 Introduction

Human actions can be represented by a serials of pose, generating images of persons with specific pose can help us more precisely construct human actions, which is helpful in the field of producing animation, games and designing actions for actors and models. Images generated by this work can also be used in generating more realistic images of human, augmenting the datasets related to human pose learning and giving the technical support to pose editing tasks. Some method use GANsYoon et al. [2020], flow fieldsLiu et al. [2020] and classic SMPLLoper et al. [2023] model to change the human pose of input image and produce images of people with proper new pose. However, those method do not have the strong ability of generalization and often suffer from keeping the basic feature of original person such as the face appearance.

In order to generate images of arbitrary person whose pose matches the given key points, we plan to leverage the knowledge of large diffusion models. Diffusion models have been widely used to generate high quality images in recent years. Large text-to-image generation models such as StableDiffusionRombach et al. [2022] and DALL-ERamesh et al. [2022] conditioned on the features of given text descriptions, gradually denoise the random images sampled from Guassion space to generate fine images that highly correlated to the input texts. However, those universal large generation models can not directly generate images with precise pattern or features such as generating

*Equal Contributors

Preprint. Under review.

images of person with specific human pose. One reason is that precise pattern information is hard to be represented by texts, the training sets of large diffusion models contain few aligned text-image samples that the text describes a precise image pattern consistent with the image. So in denoising process, given text feature as condition, diffusion model tends to map Guassion distribution to the image space that closest to the image space aligned with the conditioned text feature. ControlNet gives us a good paradigm of controlling the specific pattern of generated image. Adding a light side branch to the original diffusion model, we can inject more precise control information to the diffusion model, leading the output image distribution strictly restricted by the representations of human pose. We will combine both key points and depth map to improve the quality of generated images. This project can not only help in the human pose field, but also show how to control Diffusion generate images with fixed patterns.

2 Related work

2.1 Diffusion models

Diffusion models [Sohl-Dickstein et al., 2015] are emerging generative models that have achieved great success in image generation task [Ho et al., 2020]. In diffusion models, samples are created through a Markov chain, which begins with a random Gaussian noise and gradually denoises it into an image. The generative Markov chain is derived by reversing a forward diffusion process, which progressively transforms an image into noise. To speed up the generation process and reduce the computational requirement, Latent Diffusion Model [Rombach et al., 2022] was proposed, which compresses the image into lower-resolution latent space with a pre-trained autoencoder [Kingma and Welling, 2013, Van Den Oord et al., 2017] and then conducts the denoising process on the latent space. Stable Diffusion Rombach et al. [2022] is a large-scale implementation of latent diffusion. Based on stable diffusion, a lot of applications are developed, such as ControlNet [Zhang et al., 2023], Editanything [Gao et al., 2023] and so on.

2.2 Pose control

Regarding pose control, there are several methods commonly used at present. There are some related works as follows.

PoCoLD [Han et al., 2023] a new technique for creating controllable person images. It improves on past methods by incorporating DensePose for detailed body structure and a pose-constrained attention module for precise synthesis. PoCoLD works in latent space to enhance speed and reduce memory use, surpassing current models in image quality, speed, and efficiency. It enables pose-based image creation and appearance transfer without needing modifications to its structure.

DiffBody [Okuyama et al., 2024] introduces a one-shot method for significant edits while maintaining identity, by fitting a 3D body model to the input image, then adjusting the model's pose and shape. Initial artifacts from occlusion and shape inaccuracies are corrected through diffusion-based refinement. An iterative process using weak noise refines the whole body and face, improving structure and identity preservation. Realism is further enhanced by fine-tuning text embeddings through self-supervised learning.

2.3 ControlNet

ControlNet [Zhang et al., 2023] directly uses Stable Diffusion's own Unet as the image encoder. The network structure of ControlNet is a mirror of the Unet, and the pre-trained Stable Diffusion Unet (part) is copied as the encoder of the conditional image information, but the Decoder part is removed. Only the Encoder part and the Middle part are included, and each block is fitted with a 1×1 Convolution layer. The output of each ControlNet block is connected via a jumper to the decoder corresponding to the Unet on the right side of the Stable Diffusion.Additional requirements for image generation can be specified precisely according to user requirements, such as the pose control of the people in the image we want to implement here.

Compare with text prompts, key points and depth maps are more precise representations of human pose. In the task of generating images of person with fixed pose, images of key points and depth maps can be the additional condition off diffusion model. Given the additional condition c_f, the

target image distribution can be represent as $P(x, z|c, c_f) = P(x|z)P(z|c, c_f)$. Given c and c_f, we can first encode two conditions to get the latent variable z and then decode z to get the predicted noise of x_t

3 Problem Statement

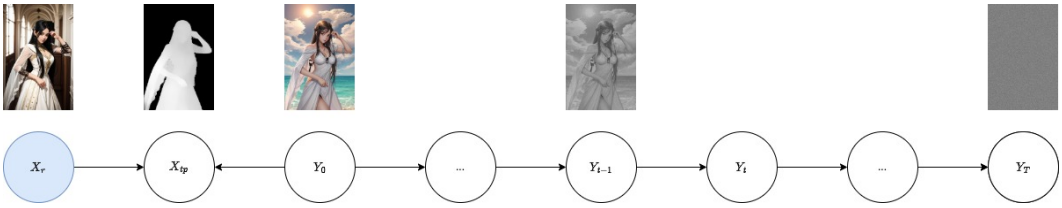


Figure 1: The PGM of the task.

In this work, we focus on the image generation task with controllable human pose. Formally, we aim to train a conditional generative model $p(y|x_r, x_t)$, which takes a reference image x_r containing the target human pose and an optional text prompt as input. The model is expected to generate a final output image y that matches the target pose (denoted as x_{tp}) in reference image x_r and also matches the text prompt if any. We present the PGM of our task in Figure 1, where X_r is observed, and we use diffusion model as the generator to predict the distribution of Y_0.

4 Methodology

Note that the reference image x_r may contains some irrelevant information, such as the background, however, we only care about the human pose and all other information should be regarded as noise. In this sense, it's practical to take a two-stages framework to accomplish the task. In the first stage, we need to build a model $x_{tp} = \mathcal{F}_\phi(x_r)$ (parameterized by ϕ) to estimate the most likely target pose x_{tp} from the reference image x_r. And then, in the second stage, we use another model $p_\theta(y|x_{tp}, x_t)$ (parameterized by θ) to approximate $p(y|x_r, x_t)$. Under this framework, the task can be modeled as $p(y|x_r, x_t) \approx p_\theta(y|\mathcal{F}_\phi(x_r), x_t)$, and the remaining question is how to choose appropriate $\mathcal{F}_\phi$ and p_θ.

4.1 Proposed Method

In this section, we introduce two novel methods aimed at enhancing posture control and image generation by leveraging segmented depth map and depth-annotated pose data. Our approaches specifically address the shortcomings of current depth-based and OpenPose-based methods. The first method improves upon existing depth-based techniques by refining the generation of image backgrounds, a common issue in such approaches. The second method enhances the fidelity of hand representations in generated images, addressing a well-known limitation of existing OpenPose-based methods.

4.1.1 Segmented Depth Map + Fine-tuned ControllNet

To advance beyond the limitations of existing depth-based methods, our first approach employs *segmented depth map* [Ranftl et al., 2020, Yang et al., 2024] as our pose estimator $\mathcal{F}_\phi$, and use *fine-tuned ControlNet* [Zhang et al., 2023] as the image generator p_θ. Compared to OpenPose [Cao et al., 2017], depth map has more semantic information about the human pose, however, it usually contains some irrelevant information like background. To tackle this issue, we use segmentation [Kirillov et al., 2023] to filter out the irrelevant information and only keep the human part in depth map (shown in Figure 2). Additionally, we find that existing ControlNet performs badly in cases where human body is partially covered by other objects. So we fine-tune the ControlNet conditioned on the segmented depth map to better fit our method.

(a) Input reference image.

(b) Original Full Depth map

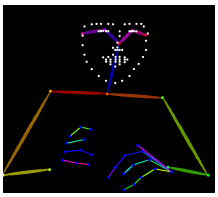
(c) Depth map masked by segmentation.

Figure 2: Comparison between original depth map and segmented depth map.

4.1.2 Annotated OpenPose + Fine-tuned ControllNet

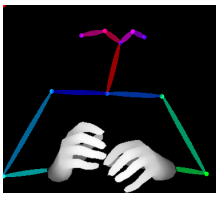
Our second method seeks to rectify the inaccuracies in hand representation found in existing OpenPose-based models. Emperically, we find that the generated fingers are usually distorted in existing OpenPose-based methods. To generate images with better fidelity, we explore to combine the advantages of pose images and depth images. The specific approach is to train a model to simultaneously control the overall pose of the character according to the pose image, and control the fine drawing of the hand with the depth image. The specific process is shown in the figure below.

(a) Raw input picture example



(b) The pose recognition result

(c) depth picture of the hand after cutting



(d) The merged final control image

We first get the pose image of the input image and the depth image of the hand respectively. Then we deleted the key points related to the knuckles of the person's hand in the pose picture, and connected the wrist to the depth image of the hand. Finally, the images are combined as control conditions, and a new controlnet model is trained to control image generation.

4.2 Alternative Methods

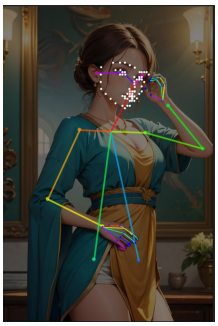
In this section, we discuss some alternative choices of $\mathcal{F}_\phi$ and p_θ and their limitations.

4.2.1 OpenPose + Existing ControllNet

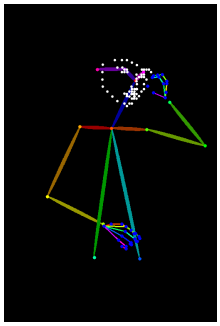
OpenPose [Cao et al., 2017] is typically used in community to estimate the human pose, which uses a set of key points representing the locations of human body joints. We can use OpenPose as pose estimator and existing ControlNet model as generator. We show how it works in Figure 4a. However, we found that it suffers from two issues. Firstly, due to the location information does not contain the orientation of the person, it is difficult to correctly identify whether the pose character is facing or facing away from the audience. For example, based on the location of key points in Figure 4b, an error example with wrong orientation is shown in Figure 4c. Additionally, due to the limited expressiveness of key points, some details may be incorrectly generated, as shown in Figure 4d, the target pose is standing while the generated pose is sitting.

4.2.2 Depth Map + Existing ControllNet

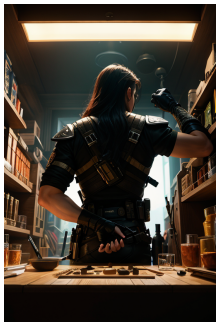
Utilizing a standalone depth map [Ranftl et al., 2020, Yang et al., 2024] as the pose estimator in conjunction with the existing ControlNet model as the generator presents certain limitations. While depth maps provide detailed information on human pose, they also capture extraneous details, including irrelevant background elements. This often results in the generated image's background



(a) Target pose recognized by OpenPose.



(b) Key points of the target pose.



(c) Generation with wrong orientation.

(d) Generation with wrong details.

Figure 4: Failure examples of estimating target pose by OpenPose, and generation by ControlNet.

being overly influenced by the depth map of the original image, leading to undesirable effects and limiting the diversity of generated images. These issues and their limitations on image generation diversity will be further explored in the results section.

4.2.3 Segmented Depth Map + Existing ControlNet

Although existing depth-based ControlNet models support the use of segmented depth maps, our experiments have identified significant limitations with this approach. The segmentation process tends to produce images with overly simplistic, pure backgrounds, lacking the complexity and realism of unsegmented approaches. While this issue can be somewhat mitigated through the use of carefully crafted prompts, a more pressing concern arises when the subject is obscured by an object. In such cases, the segmentation process inadvertently removes the obstructing object, leading to inaccuracies in the representation of the subject. These challenges, along with illustrative examples, will be discussed in detail in the results section.

5 Experiment

5.1 Implementation details

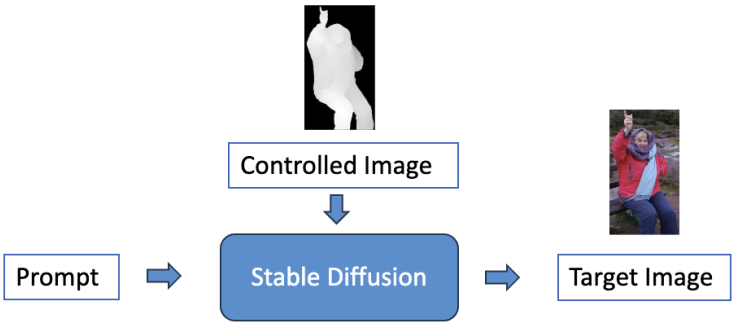


Figure 5: The training approach.

In this section, we put details in the training process and dataset to fine-tune our model.

Training Approach Figure 5 presents the three essential elements of our training approach: a controlled image to direct the model's output, a textual prompt for semantic guidance, and a target image that sets the generation objective. Together, these components enable the Stable Diffusion model to learn and replicate the desired outputs during training.

Dataset In the initial phase of our study, we employed the LIP dataset for training. However, due to insufficient images, the results derived from this dataset alone were not satisfactory in terms of model performance. To enhance the training outcomes and improve model accuracy, we subsequently integrated an additional dataset, the COCO dataset Lin et al. [2015], into our training regimen. The COCO2017 contains 64115 real-world images categorized as humans for training and 2693 human images for validation. All images in training set and validation set are having annotations which contain the bounding box and rough segmentation area of target persons. To ensure the target persons in training images have high resolution and fine appearance, we first cut the training image by using their bounding boxes and delete the images whose sizes are smaller than 10000. Then we segment images by using lang-segment-anything given prompt "person" to get the mask of target person. We use l2 distance to compute the difference d between mask generated by lang-segment-anything and rough segmentation area. For each image, if d is under threshold 0.01, predicted mask of target person matches ground truth with little background noise. We delete images whose d is greater than threshold. After filtering process, we get 11070 images for training and 617 images for testing.

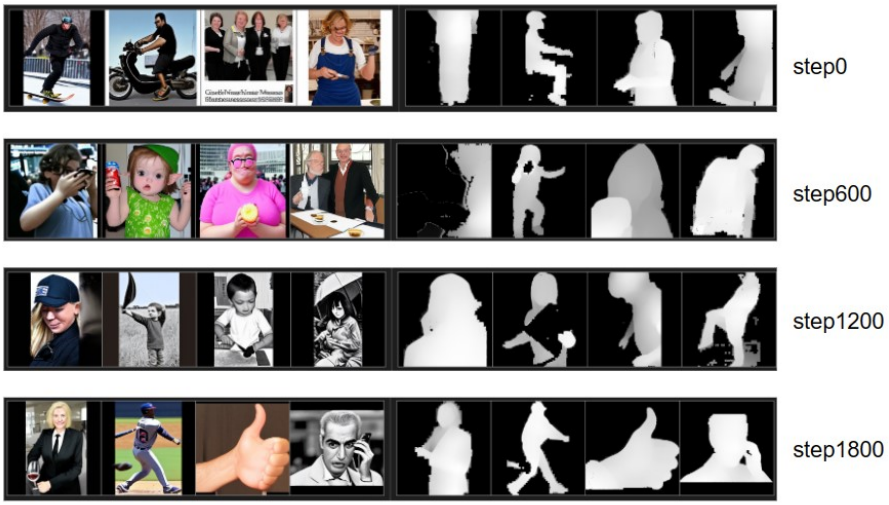


Figure 6: The images sampled during training

Training Data for the Segmented Depth Model The training of the segmented depth model involved three types of input data:

1. *Depth Map of the Original Image:* Generated through the application of the MiDaS depth estimation technique.
2. *Background-Masked Image:* Created by employing the "segment anything" library for background segmentation.
3. *Image-Specific Prompts:* Generated using the CLIP (Contrastive Language-Image Pre-Training) model developed by OpenAI, which provided contextual prompts for each image.

Both 1 and 2 are then merged as a single controlled image during training.

Training Data for the Depth-Annotated Pose Model The depth-annotated pose model was trained using a distinct set of inputs:

1. *Depth Map of the Hand:* The hand's depth was ascertained using the MeshGraphormer depth reconstruction library.
2. *Posture Points of the Image:* The posture data was extracted from the original image using the OpenPose library, which identifies human pose.
3. *Image-Specific Prompts:* Similar to the segmented depth model, the CLIP model from OpenAI was utilized to create relevant prompts for each image.

Similarly, Both 1 and 2 are then merged as a single controlled image during training.

These datasets and inputs were meticulously chosen to ensure the comprehensive training of our models, aiming for high accuracy in segmented depth analysis and depth-annotated pose estimation.

5.2 Training Process

Figure 6 shows the images sampled from DDIM process in different training steps. In first training steps, persons in generated images are having arbitrary poses. After about 1000 steps of training, parameters of side UNet start to control poses consistent with conditional depth maps.

5.3 Results of Segmented Depth Map

(a) Original depth map of reference image.

(b) An image based on original depth map.

(c) Segmented depth map.

(d) An image based on segmented depth map.

Figure 7: Examples of image generatation using depth.

(a) Input reference image.

(b) Segmented depth map.

(c) Image generated by existing ControlNet.

(d) Image generated by our model.

Figure 8: Comparison between existing ControllNet and our fine-tuned model, contidionted on segmented depth map.

In this section we will show that our segmented depth's advantage over existing methods. The existing methods employing a standalone depth map can indeed manipulate the pose of the target figure. Nevertheless, these depth maps capture excessive detail, encompassing not only the subject's pose but also extraneous elements. This is evident in Figures 7b, where the backgrounds of the generated images are significantly influenced by the depth map (see Figure 7a,the depth of the background is also recognized and applied to image generation), leading to undesirable effects on the output images and reducing their variety.

As depicted in Figure 7d(based on Figure 7c), our segmented depth methodology adeptly manages to maintain the posture of the target image while allowing for customizable backgrounds.

The existing models falter in scenarios where the subject is partially obscured by objects(see Figure 8b). The deficiency of this approach in handling occlusions is demonstrated in the Figure 8c. However, our model demonstrates robust performance even when the subject is obscured by an object. The subsequent figure illustrates our model's capability to accurately render the target image, even when the individual in the source image is obscured by an object, such as a book. See Figure 8d.

5.4 Results of Depth Annotated Pose

Figure 9 showcases an example from our Depth Annotated Pose model, illustrating its capability to accurately generate images with the desired posture, where the hands and arms are seamlessly connected, unlike the outcomes from existing approaches.

Current methods for simulating human posture with accurate hand shapes typically rely on a two-step process: a posture control model is applied first, followed by a separate depth model for the hands. This approach often results in a noticeable disconnection between the hands and arms, a consequence of the two models being developed and trained independently, which can lead to inconsistencies when they are used together. Figure 9c displays the results of such a method, clearly showing the unnatural separation between the hand and arm regions.

(a) Input reference image. (b) Depth annotated pose image. (c) Image generated by existing ControlNet. (d) Image generated by our model.

Figure 9: Examples of depth annotated pose generated by our fine-tuned model.

Performance In our tests on a 3080ti graphics card for 512x512 pixel images, using existing OpenPose and depth models consumed 3.88GB of memory and took 3.2 seconds per image. Switching to our custom-trained model reduced memory usage to 2.7GB and cut rendering time to 2.6 seconds, showcasing our model's superior efficiency and performance.

6 Conclusion

In this paper, we present PostureCraft to control human pose generation with diffusion models given an reference image. We propose segmented depth map control, where we masked the depth map by the segmentation to filter out irrelevant noise. To overcome the distortion in generated human hands/fingers, which is a well-known issue of human image generation, we propose depth-annotated OpenPose to provide a better representation of human hands. We fine-tuned the existing ControlNet to better fit our proposed methods. The experimental results show that our methods can precisely control the generated human pose and present accurate human hands, which proves the effectiveness of our proposed methods.

References

Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017.

Shanghua Gao, Zhijie Lin, Xingyu Xie, Pan Zhou, Ming-Ming Cheng, and Shuicheng Yan. Editanything: Empowering unparalleled flexibility in image editing and generation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 9414–9416, 2023.

Xiao Han, Xiatian Zhu, Jiankang Deng, Yi-Zhe Song, and Tao Xiang. Controllable person image synthesis with pose-constrained latent diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22768–22777, 2023.

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

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015.

Wen Liu, Zhixin Piao, Zhi Tu, Wenhan Luo, Lin Ma, and Shenghua Gao. Liquid warping GAN with attention: A unified framework for human image synthesis. *CoRR*, abs/2011.09055, 2020. URL `https://arxiv.org/abs/2011.09055`.

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. *SMPL: A Skinned Multi-Person Linear Model*. Association for Computing Machinery, New York, NY, USA, 1 edition, 2023. ISBN 9798400708978.

Yuta Okuyama, Yuki Endo, and Yoshihiro Kanamori. Diffbody: Diffusion-based pose and shape editing of human images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6333–6342, 2024.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022.

René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence*, 44(3):1623–1637, 2020.

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.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015.

Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. *arXiv preprint arXiv:2401.10891*, 2024.

Jae Shin Yoon, Lingjie Liu, Vladislav Golyanik, Kripasindhu Sarkar, Hyun Soo Park, and Christian Theobalt. Pose-guided human animation from a single image in the wild. *CoRR*, abs/2012.03796, 2020. URL `https://arxiv.org/abs/2012.03796`.

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.