Pose-dIVE: Pose-Diversified Augmentation for Person Re-Identification – Supplementary Materials –

Inès Hyeonsu Kim[1*] Woojeong Jin[1*] Soowon Son[1] Junyoung Seo[1] Seokju Cho[1]
JeongYeol Baek[2] Byeongwon Lee[2] JoungBin Lee[1] Seungryong Kim[1†]
KAIST AI[1] SK Telecom[2]

A. Detailed Explanation of MSMT17-S

In the domain of person re-identification (Re-ID), addressing biases within datasets is crucial for improving model generalization. Such biases can lead to models that perform well under familiar conditions but struggle when faced with new scenarios. Among the various biases present, we focused on addressing camera viewpoint bias and human pose bias inherent in the dataset through dataset augmentation. To rigorously evaluate and challenge the generalization capabilities of our Pose-dIVE method, we create two specialized versions of the MSMT17 dataset: MSMT17-S (Cam) and MSMT17-S (Pose).

MSMT17-S (Cam) is specifically designed to amplify the bias related to camera viewpoints. In this version of the MSMT17 dataset, the training set consists of images captured from a specific set of camera angles, while the test set includes entirely different angles that do not overlap with those in the training data. This deliberate split maximizes the viewpoint bias in the training data, creating a scenario where the model must generalize to new, unseen camera angles during testing. MSMT17-S (Cam) serves as a challenging benchmark to evaluate whether the Pose-dIVE augmentation method can effectively help models overcome viewpoint biases, which are prevalent in real-world Re-ID scenarios where camera placements vary widely.

MSMT17-S (Pose), on the other hand, is designed to emphasize the bias related to human poses. In this dataset, the training data is composed of individuals captured in a specific range of poses, while the test set features entirely different poses that are not present in the training data. By maximizing pose bias in this way, MSMT17-S (Pose) creates a challenging environment where a Re-ID model must generalize to new poses that it has never seen during training. This setup is particularly relevant for testing how well Pose-dIVE can assist a model in adapting to a wide variety of human postures, which is crucial for accurate person identification across different activities and positions.

These datasets specifically highlight the challenges posed by biases in camera viewpoints and human poses, respectively, allowing us to assess the effectiveness of the Pose-dIVE augmentation strategy in mitigating these biases.

B. Conditional Diffusion Models for Pose-Diversified Augmentation

In this section, we delineate the architecture of the conditional diffusion model we employed. This model is designed to generate images with diverse camera viewpoints and human poses, conditioned on the relevant input data. However, naively training the generative model on a human Re-ID dataset without careful consideration may produce degenerated results for camera viewpoints or human poses that are rarely present in the training dataset.

We address this problem by leveraging the vast knowledge in pre-trained Stable Diffusion [3]. We first provide a preliminary explanation of the Stable Diffusion model [3], followed by the method for augmenting the images in the training dataset while controlling their human pose, camera viewpoint, and identity. The overall architecture can be seen in Figure 1.

Preliminary: Stable Diffusion. Diffusion model [1, 3, 5] is a generative model that samples images from the learned data distribution $p(x)$ through iterative denoising process from Gaussian noise. Our method builds upon Stable Diffusion (SD) [3]. SD performs a denoising process in a latent space of Autoencoder [6], reducing the computational cost compared to denoising in the pixel space [1, 5]. Specifically, the encoder in SD maps a given image $\mathbf{x}$ into a latent representation $\mathbf{z}$, denoted as $\mathbf{z} = \mathcal{E}(\mathbf{x})$.

During training, SD learns a denoising U-Net [4] ϵ_θ that predicts normally distributed noise ϵ given a noised latent $\mathbf{z}_t$, which is a noisy latent of $\mathbf{z}$ with a Gaussian noise at noise level t. This U-Net function can be trained with a following objective:

$$\mathcal{L} = \mathbb{E}_{\mathbf{z}_t, c, \epsilon, t} \left(\| \epsilon - \epsilon_\theta \left(\mathbf{z}_t, c, t \right) \|_2^2 \right), \tag{1}$$

where c denotes conditional information for generation. The condition, which is a text prompt encoded using the

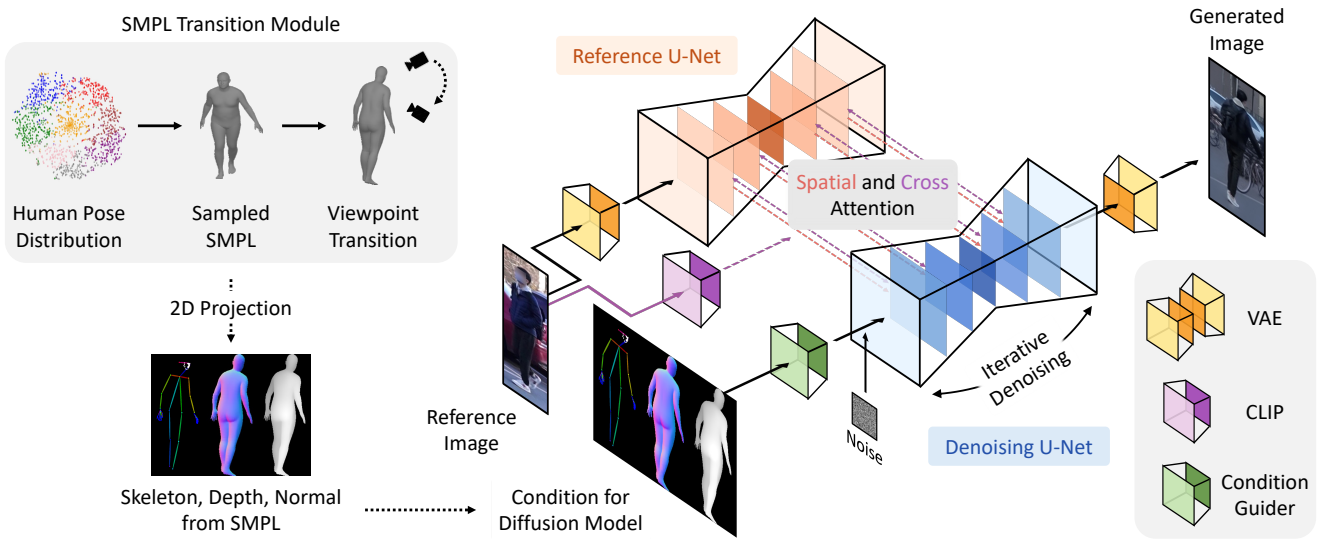

Figure 1. **Overall architecture of generative model in Pose-dIVE.** Given the viewpoint and pose distributions, we first render the body shape sampled from the distribution using SMPL, generating the corresponding skeleton, depth map, and normal maps. These conditions, along with a reference image for identity preservation, are then fed into generative module, which consists of two branches: the reference U-Net processes the identity information from the reference image, while the denoising U-Net generates a person with the same identity, given the input conditions. The denoising U-Net generates images by iterating through the denoising process.

CLIP text encoder [2], enables controllability over the image generation process. The denoising U-Net is composed of three parts: downsampling block, bottleneck block, and upsampling block. Each block consists of a combination of 2D convolutional layers, self-attention layers, and cross-attention layers.

During inference, a sample $\mathbf{z}_T$ from a Gaussian distribution is gradually denoised using the trained denoising U-Net. Undergoing the denoising process from $t = T$ to $t = 0$, the model generates $\mathbf{z}_0$. This final latent representation is then passed through the decoder $\mathcal{D}$ to produce the output image.

Injecting human pose and camera viewpoint into diffusion model. For conditions that should be spatially aligned with the generated output, we process them with a respective pose guider network and concatenate the processed conditions along the channel dimension. Specifically, the depth map $\mathbf{d} \in \mathbb{R}^{H\times W\times 1}$, surface normals $\mathbf{n} \in \mathbb{R}^{H\times W\times 3}$, and rendered human skeleton $\mathbf{s} \in \mathbb{R}^{H\times W\times 3}$ are each processed by a respective pose guider network. This network reduces the spatial size of the condition to $1/8$ of the original size and embeds the pose information into $\mathbb{R}^{\frac{H}{8}\times\frac{W}{8}\times C}$ embeddings, aligning it with the size of the latent representation in the diffusion model. The last layer of the pose guider network is initialized to zeros to minimize the initial degradation during the fine-tuning stage of the pre-trained SD model. The processed conditions are then concatenated along the channel dimension and added to the projected noise before it is fed into the U-Net.

Injecting identity into diffusion model as a condition. For human identity, unlike the condition that is spatially aligned with the output, it is not necessarily aligned with the output. In this regard, instead of adding the condition pixel-wise, we provide the identity information to the denoising U-Net with attention. Specifically, we design a reference U-Net that has the same architecture as the denoising U-Net, while the weights of the U-Net are initialized with a pre-trained Stable Diffusion model. To inject the identity of an image into the denoising U-Net, we first feed the image into the reference U-Net. Then, the identity information is shared with the denoising U-Net using self-attention for each block. In more detail, given the intermediate feature map from the denoising U-Net $f_1 \in \mathbb{R}^{(h\times w)\times c}$ and from the reference U-Net $f_2 \in \mathbb{R}^{(h\times w)\times c}$, they are concatenated along the spatial dimension, followed by the self-attention layer. Then, the first half of the output is used as the input for the following layers in the denoising U-Net. In this way, the two parallel branches can benefit from the extensive pre-trained knowledge of Stable Diffusion. Additionally, the identical architecture of the two branches facilitates training by sharing the same feature space. For the cross-attention part in Stable Diffusion where text embeddings from CLIP are used, we instead utilize image embeddings from the CLIP image encoder. This is possible because both text and image embeddings are trained to reside in the same embedding space.

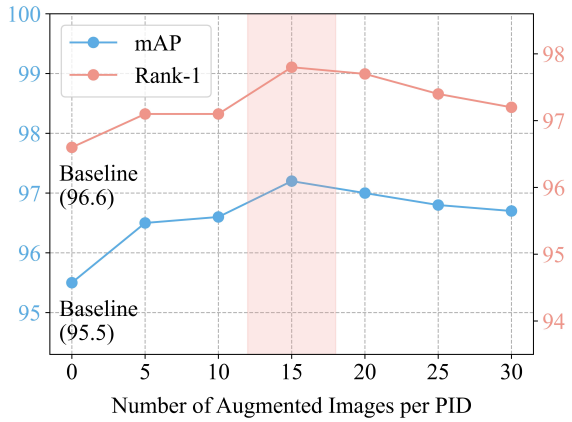

Figure 2. **Impact of the number of generated images per PID.** Experiments are conducted in the Pose-dIVE augmented CUHK03 (L) dataset. We use CLIP-ReID baseline.

C. Additional qualitative results

In Figure 3, we provide additional qualitative examples from the MSMT17, Market-1501, CUHK03 (D), and CUHK03 (L) datasets. The generated images maintain a high level of realism, effectively capturing the nuances of natural human appearances. Also, they successfully preserve the identity of individuals from the input reference images, ensuring that the augmented data remains faithful to the original person's characteristics. This is particularly important for re-identification tasks where identity preservation is critical. These additional results further validate the effectiveness of our proposed augmentation method across different datasets.

D. Analysis on the number of generated images

Figure 2 illustrates the impact of varying the number of generated images on the performance of the Re-ID model when using our proposed augmentation strategy. To assess this, we progressively increased the number of generated images in the training dataset and trained the CLIP-reID model, carefully monitoring performance changes at each increment. Note that this augmentation was applied only to the training dataset, while the test dataset remained unchanged throughout the experiments. Our findings indicate that generating approximately 15 images per person yields the highest performance for the Re-ID model.

E. Filtering Protocols for High-Quality Augmentation

To guarantee high-quality outputs in our augmented dataset, we applied a series of filtering procedures during both the generative model training and post-processing steps. Our filtering method uses pose scores obtained from the human pose estimation process, retaining only images that exceed certain thresholds. Filtering is applied to three types of images: *reference images*, *target images*, and *generated images*.

Let $\mathcal{D} = \{(I^i_{\text{ref}}, I^i_{\text{tar}})\}_{i=1}^N$ denote the data set, where $I_{\text{ref}}, I_{\text{tar}} \in \mathbb{R}^{H\times W\times C}$ represent the reference image and target image, respectively. The reference image provides the identity for the generated image, and the target image specifies the desired pose. The pose estimation model $\mathcal{P} : \mathbb{R}^{H\times W\times C} \rightarrow \mathbb{R}^K$ outputs confidence scores for K keypoints:

$$\mathcal{P}(\mathcal{I}) = \{\sigma_k(I)\}_{k=1}^K, \tag{2}$$

where $\sigma_k(I) \in [0, 1]$ is the confidence score for keypoint $k \in \mathcal{K}$ corresponding to each body joint.

Filtering Protocol. We use predefined thresholds τ to filter out the images that do not meet our quality criteria. First, we remove input images with heavy occlusions. For instance, when the lower body is significantly occluded, generating a full-body image from such an input becomes an ill-posed problem. Formally, we retain I_{ref} if:

$$\min(\{\sigma_k(I_{\text{ref}})\}_{k\in\mathcal{K}}) \geq \tau_{\text{ref}}, \tag{3}$$

where τ_{ref} is the threshold for the reference image. Next, we also filter out target pose images with notable occlusions, prioritizing training samples where the entire body is clearly visible. This prevents cases where the input image is clear, but the generated output unexpectedly shows occlusions. We retain I_{tar} if:

$$\frac{1}{N_k}\sum_{k=1}^{K}\sigma_k(I_{\text{tar}}) \geq \tau_{\text{tar}}, \tag{4}$$

where τ_{tar} is the threshold for the target image. After generating augmented images, we apply a final filtering step to discard outputs that do not align with desired poses. This ensures that the generated images are of high quality and consistent with the target pose. We first compute the mean absolute difference in confidence scores between the generated and target poses:

$$\mathcal{O} = \frac{1}{N_k}\sum_{k=1}^{K}|\sigma_k(I_{\text{gen}}) - \sigma_k(I_{\text{tar}})|. \tag{5}$$

To ensure the generated image maintains a high overall keypoint confidence score, we use I_{gen} only if both conditions are satisfied:

$$\mathcal{O} \leq \epsilon_{\text{gen}} \quad \text{and} \quad \frac{1}{K}\sum_{k=1}^{K}\sigma_k(I_{\text{gen}}) \geq \tau_{\text{gen}}, \tag{6}$$

where ϵ_{gen} is a small value that determines whether the poses between the generated image and the target image are sufficiently aligned.

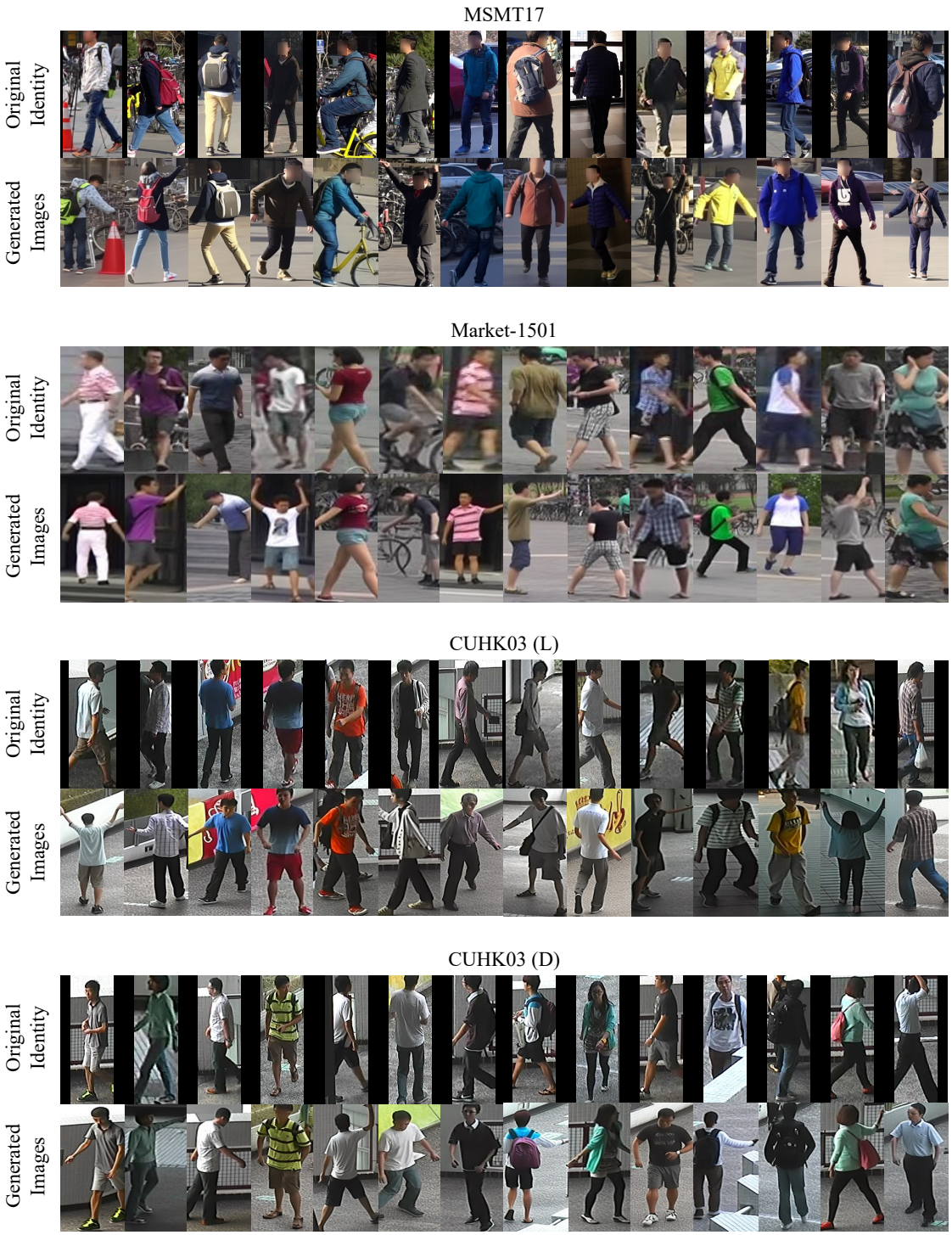

Figure 3. **Additional qualitative results.** Examples of generated images from the Pose-dIVE augmented datasets. The results demonstrate realistic rendering while preserving the identity of the reference images.

Through these three filtering stages, we ensure a more consistent and higher-quality augmented dataset for Re-ID training. The use of confidence scores σ_k directly from the pose estimation model allows for effective and interpretable quality control.

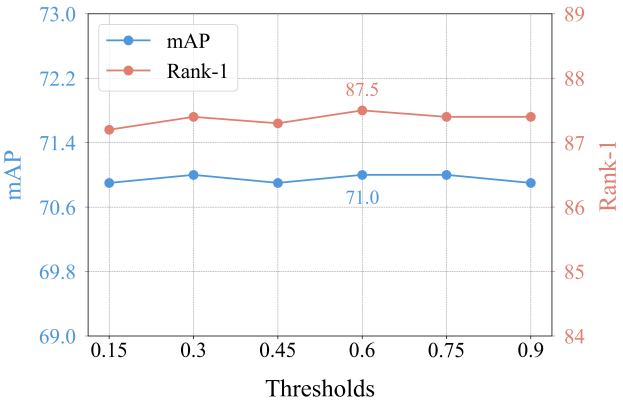

Figure 4. **Ablation study of pose thresholds for filtering generated data.** We matched the number of training samples to ensure fairness.

F. Analysis on Filtering Threshold Selection

We analyzed the sensitivity of our filtering protocol to the pose confidence threshold. By varying the threshold from 0.15 to 0.9, we evaluated the CLIP-reID model on the MSMT17 dataset. Our results, presented in Figure 4, show that performance remains stable across this broad range. The mAP and Rank-1 accuracy curves are consistently high, peaking at a threshold of 0.6. This indicates that our filtering methodology is robust and not highly sensitive to the chosen threshold. We hypothesize that as long as the generated image preserves the reference identity, it provides a valuable supervision signal, even if its pose is slightly degraded.

G. Failure Cases and Discussions

We analyzed around 100 samples in our augmented datasets and identified two failure cases. The first involves scenarios where the input image does not show a backpack, yet the generated image does. Although this could be considered a failure for a straightforward image-to-image generation task, it is less problematic for person re-identification because a person in the real world can appear both with and without a backpack. Generating such "hard" samples can rather actually benefit the Re-ID model by enhancing its ability to handle variations in appearance. The second failure case arises when the input images have poor quality, such as noisy or very low-fidelity images. We believe that tackling these problems will require additional future work.

H. Visualizing Conditions in Generative Models

We present visualizations of conditions used in our generative model, specifically, the skeleton, depth, and normal maps, in Figure 5. With these conditions, we can control the human pose and viewpoint of the images generated by our diffusion models.

I. Controlled Qualitative Comparisons

To complement the qualitative results in the main paper, we provide additional controlled comparisons in Figure 6. These results are designed to explicitly evaluate identity preservation and pose/viewpoint disentanglement under two conditions. First, we show multiple poses and viewpoints for the same identity in Figure 6 (a). Despite large spatial changes, our method consistently preserves identity-specific cues such as clothing and accessories. Second, we fix the same pose and viewpoint across different identities in Figure 6 (b). This setting highlights that our approach can generate diverse individuals while maintaining consistency in pose and viewpoint. Together, these comparisons provide a more systematic validation that Pose-dIVE effectively separates pose/viewpoint variations from identity, enabling robust person re-identification.

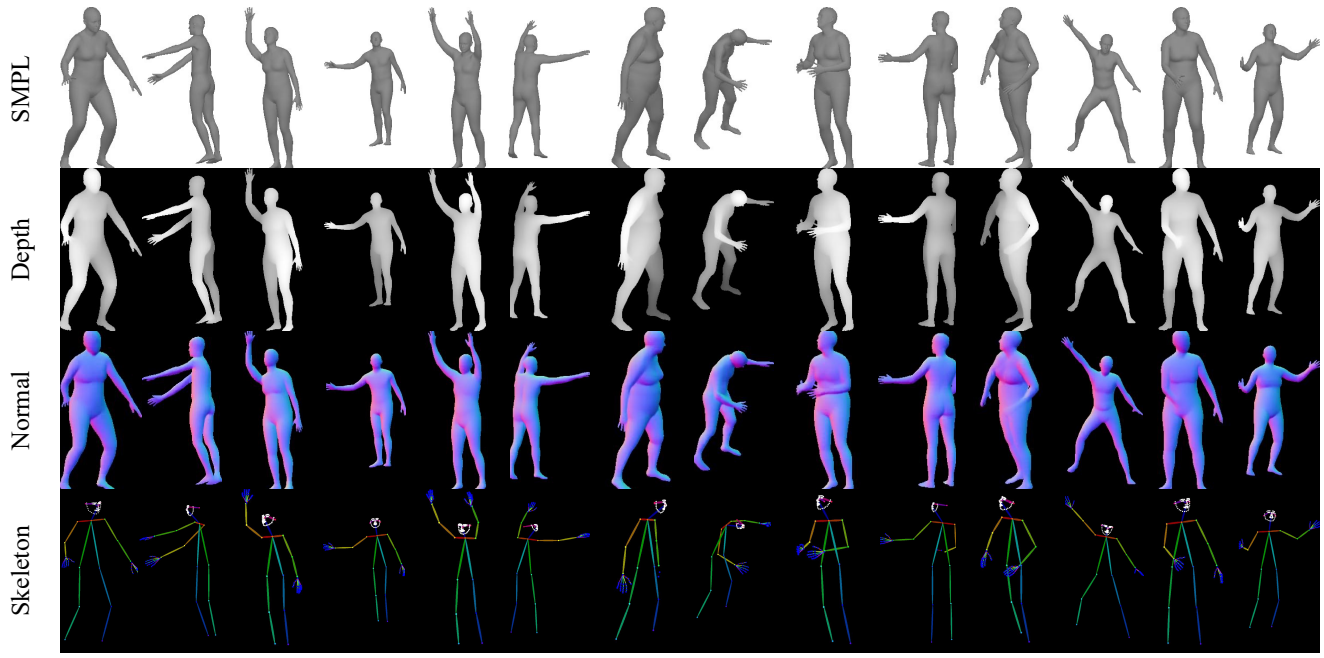

Figure 5. **Example SMPL, skeleton, depth and normal maps from external dataset.** Examples of generated images from the PosedIVE augmented datasets. The results demonstrate realistic rendering while preserving the identity of the reference images and aligning accurately with the target poses.

(a) Same pose with different identities (b) Same identity with different poses

Figure 6. **Controlled qualitative comparisons.** (a) Different identities under the same pose and viewpoint. Identity diversity is maintained despite identical pose/viewpoint conditions. (b) Same identity across different poses and viewpoints. Identity cues are faithfully preserved while accommodating significant spatial changes.

References

[1] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 1

[2] 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. 2

[3] 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. 1

[4] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015. 1

[5] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 1

[6] Jiahui Yu, Xin Li, Jing Yu Koh, Han Zhang, Ruoming Pang, James Qin, Alexander Ku, Yuanzhong Xu, Jason Baldridge, and Yonghui Wu. Vector-quantized image modeling with improved vqgan. *arXiv preprint arXiv:2110.04627*, 2021. 1