Controllable Person Image Synthesis with Pose-Constrained Latent Diffusion

Xiao Han[1,2] Xiatian Zhu[1,4] Jiankang Deng[3] Yi-Zhe Song[1,2] Tao Xiang[1,2]
[1] CVSSP, University of Surrey [2] iFlyTek-Surrey Joint Research Centre on AI
[3] Imperial College London [4] Surrey Institute for People-Centred AI
{xiao.han, xiatian.zhu, y.song, t.xiang}@surrey.ac.uk j.deng16@imperial.ac.uk

Abstract

Controllable person image synthesis aims at rendering a source image based on user-specified changes in body pose or appearance. Prior art approaches leverage pixel-level denoising diffusion models conditioned on the coarse skeleton via cross-attention. This leads to two limitations: low efficiency and inaccurate condition information. To address both issues, a novel ***Po****se-****Co****nstrained* ***L****atent* ***D****iffusion model (****PoCoLD****) is introduced. Rather than using the skeleton as a sparse pose representation, we exploit DensePose which offers much richer body structure information. To effectively capitalize DensePose at a low cost, we propose an efficient pose-constrained attention module that is capable of modeling the complex interplay between appearance and pose. Extensive experiments show that our PoCoLD outperforms the state-of-the-art competitors in image synthesis fidelity. Critically, it runs* $2\times$ *faster and consumes* $3.6\times$ *smaller memory than the latest diffusion-model-based alternative during inference. Our code and models are available at https://github.com/BrandonHanx/PoCoLD.*

1. Introduction

The task of Controllable Person Image Synthesis (CPIS) is to modify a source image according to the user-specified changes in body pose or appearance [2, 23, 30]. Underpinning a wide variety of applications in virtual and augmented reality, gaming, and fashion [11, 12], there has been increasing attention in the computer vision community. In particular, modeling the large pose deformations in the 2D appearance caused by 3D pose changes is one of the biggest changes in CPIS [18]. This is further compounded by the inevitable complex self-occlusion of the human body, causing further uncertainties in predicting unobserved appearance for the target pose. Consequently, having the generative CPIS model understand the whole image *contextually* is indispensable in order to achieve plausible synthesis.

Generative Adversarial Networks (GAN) [9, 26] have been the major architecture used in CPIS [25, 30, 31, 35].

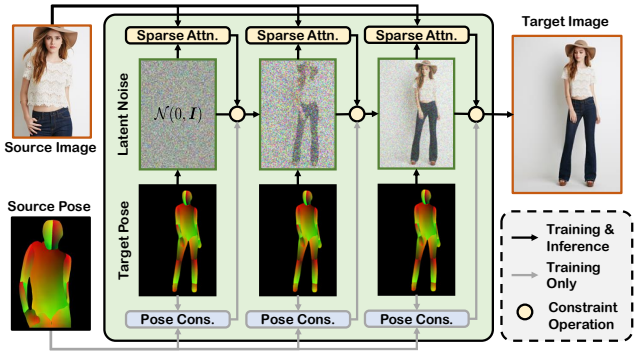

Figure 1. Our PoCoLD is a *latent diffusion model* that can handle person image synthesis with pose or appearance control. It is conditioned on the target DensePose map and the appearance of the source image. An efficient *pose-constrained attention* is proposed to explicitly regularize the denoising learning process.

However, these existing models are challenged by the need of preserving a consistent body structure and garment texture in a single forward pass. Recently, Diffusion Models (DM) [14, 36] have emerged as a favorable alternative to GAN, additionally with more stable optimization and simpler loss function design. Inspired by the massive advance of image generation and editing [6, 29, 32, 33], diffusion models have been recently exploited for CPIS with the best-ever results achieved [2]. However, this diffusion-based method comes with a couple of drawbacks: (1) ***Mismatch between the sparse pose condition (i.e., body skeleton) and the source image with dense details***: When modeling their interaction, such information intensity mismatch might lead to ill association, finally hurting the final synthesis. (2) ***Slow inference*** as the denoising diffusion takes place in the high-resolution pixel space iteratively. For instance, prior diffusion-based art [2] needs approximately 10 seconds for one 256×176 image generation on a machine with one V100 GPU.

To address the aforementioned limitations, a novel **Po**se-**Co**nstrained **L**atent **D**iffusion (**PoCoLD**) method is pro-

posed in this work. Although latent diffusion process [32] is much more efficient than pixel-level diffusion, adopting it for CPIS is non-trivial because once the source image is converted into a latent feature space, the original structural information is prone to be distorted during conditional denoising. To alleviate this problem, PoCoLD is formulated by using DensePose [10] as the pose control with much denser structural pose and shape information compared to body skeleton. Moreover, we design a *pose-constrained attention module* to effectively and efficiently integrate the source image information with the target pose (Fig. 1). In particular, the pose constraint is derived from both the source and target DensePose maps for calibrating the sparsified attention prediction. Consequently, the appearance details of the source image with correspondence to local regions of the target pose can be better captured, leading to more accurate and realistic synthesis (Fig. 5).

Our ***contributions*** are as follows: (1) We propose the first latent diffusion-based method for controllable person image synthesis with condition on DensePose based user control. (2) A new pose-constrained attention module is formulated for effectively and efficiently modeling the non-trivial interplay between source appearance and target pose. (3) Extensive experiments on the DeepFashion [20] benchmark show that our PoCoLD sets new state of the art under the key performance metrics (*e.g.*, SSIM and LPIPS) whilst enjoying significantly faster inference than prior diffusion-model-based alternative. A user study is also conducted to further validate the superiority of our model in the quality of generated images. Also, we show that our model can be applied for more tasks (*e.g.*, pose-only conditioned person image synthesis and appearance transfer) without architectural change and further optimization.

2. Related work

Controllable person image synthesis presents a primary obstacle in capturing the intricate structure of the spatial transformation of the pose while preserving the fine-grained details of the textures. It has been studied extensively by computer vision researchers, especially with the unprecedented success of GAN-based models [9,26] for conditional image synthesis. Early attempts [8, 23, 24] utilized pose-irrelevant features to extract appearance features. However, to better represent complex textures, subsequent works attempted to use human parsing maps [5, 22, 25, 46, 49] or DensePose UV maps [34]. Despite this, the final output may still wash out detailed patterns if the generation process is modulated uniformly. To achieve spatially-adaptive modulations, flow-based methods [1, 18, 31, 35, 39] were proposed to estimate appearance flow between reference and desired targets, trained with either unsupervised method or pre-calculated labels obtained by 3D models of human bodies [21]. Later, attention-based methods [30,38,47,50] have become mainstream for this task, as they can extract dense correspondences even under complex deformations or severe occlusions. Recently, the Diffusion Model (DM) [2] has emerged as an alternative approach, which breaks down the problem into a series of forward-backward diffusion steps [6, 14] to learn plausible transfer trajectories and thus achieves promising results. As described earlier, it suffers from both effective and efficient pose control, which are addressed in our method.

Diffusion models [14, 32, 36, 37] are emerging generative models that can generate competitive or even better content than GANs. The core idea behind DMs is to start with a low-quality noise signal and gradually refine it over a series of steps to generate high-quality samples, making it suitable for high-fidelity and context-aware generations. After success in the unconditional generation [6, 16], these models are extended to work in conditional generation settings. Promising results of image generation are first reported under class-condition [15], and then DMs achieve unbelievable results under semantic map [32], exemplar [45], sketch [43], natural language [27, 32, 33], instruction [3], and so on. Notably, in order to reduce the computation power required for training and inference DMs on the pixel space, Latent Diffusion Model (LDM) [32] was recently introduced, where the denoising steps are conducted on the compressed lower-resolution latent space of a powerful pre-trained autoencoder [7, 40]. Our model also adopts an LDM-based architecture. Importantly, a novel pose-constrained attention module is introduced in this work to further improve the inference efficiency and quality.

3. Methodology

We aim to train a conditional generative model $p_\theta(\boldsymbol{y}|\boldsymbol{x}_s, \boldsymbol{x}_{tp})$, parameterized as $\boldsymbol{\epsilon}_\theta$, which takes a source image $\boldsymbol{x}_s$ and a target pose $\boldsymbol{x}_{tp}$ as input. The model is expected to generate a final output image $\boldsymbol{y}$ that matches the target pose $\boldsymbol{x}_{tp}$ and also retains the same texture in the source image $\boldsymbol{x}_s$. We adopt the latent diffusion model [32] for faster training and inference, in contrast to pixel-level diffusion [2]. We further introduce an efficient pose-constrained attention mechanism for extracting the textures of the source image more accurately during the denoising process. Next, we start with a brief overview of diffusion models for person image synthesis, followed by a detailed description of our PoCoLD in Sec. 3.2. A general overview of our method is provided in Fig. 2.

3.1. Background and preliminaries

Person image synthesis via diffusion model. A diffusion model can break down the CPIS problem into a series of forward-backward diffusion steps by learning plausible transfer trajectories [2]. The main concept behind

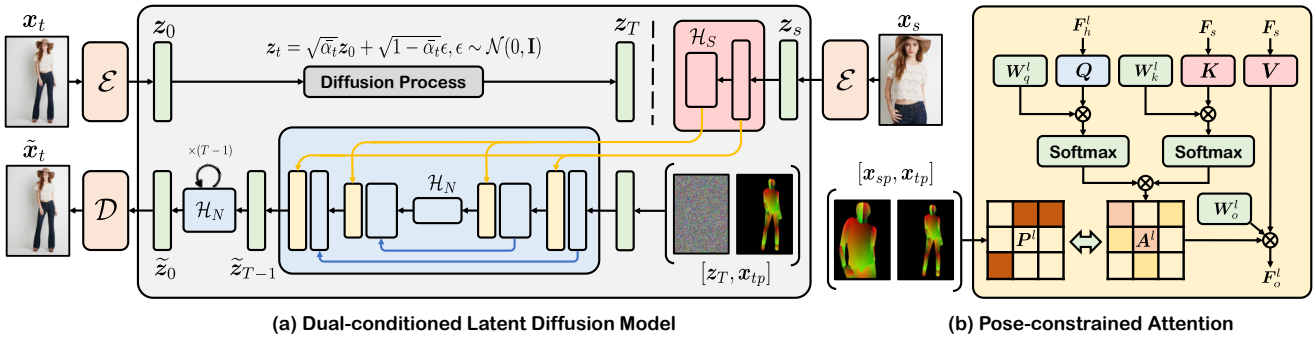

Figure 2. **Architecture overview.** Our proposed PoCoLD is **(a)** a UNet-based *latent diffusion model* composed of a *pre-trained auto-encoder* ($\mathcal{E}, \mathcal{D}$ denote the encoder and decoder respectively), a *denoising diffusion module* $\mathcal{H}_N$ and a *source image encoder* $\mathcal{H}_S$. Specifically, the image encoder $\mathcal{H}_S$ first extracts the texture patterns of a source image $\boldsymbol{x}_s$ into multi-scale feature maps. With the denoising diffusion module $\mathcal{H}_N$, the feature maps are then attended by the concatenated target pose condition $\boldsymbol{x}_{tp}$ and noisy latent $\boldsymbol{z}_t$ at every time step t for iterative synthesis during sampling. More specifically, **(b)** a novel *pose-constrained attention* module is designed to impose structural constraints derived from the geometrical relationship between the source pose $\boldsymbol{x}_{sp}$ and the target pose $\boldsymbol{x}_{tp}$. To be more efficient, the attention is sparsified by downsampling the query and key.

the diffusion model is to progressively add Gaussian noise $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I})$ to a target $\boldsymbol{y}$, resulting in a noisy image $\boldsymbol{y}_t$ at each timestep $t \in T$ [6, 14, 36]. The noise level increases over time. The reverse mapping is learned through a backward denoising process. To incorporate the target pose $\boldsymbol{x}_{tp}$ and source image $\boldsymbol{x}_s$ in the noise prediction process, a previous work [2] uses a dual conditioned architecture. Specifically, $\boldsymbol{x}_{tp}$ is represented by a pose skeleton and channel-wise concatenated with $\boldsymbol{y}_t$ at each timestep, while $\boldsymbol{x}_s$ is first encoded by an encoder and then attended by $\boldsymbol{y}_t$ through vanilla cross-attention [32,41]. The whole model is optimized to predict the added noise using a standard Mean Squared Error (MSE) loss:

$$\mathcal{L}_{\text{mse}} = \mathbb{E}_{\boldsymbol{y},\boldsymbol{x}_s,\boldsymbol{x}_{tp},\boldsymbol{\epsilon},t} \left\| \boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta \left(\boldsymbol{y}_t, t, \boldsymbol{x}_{tp}, \boldsymbol{x}_s \right) \right\|^2. \quad (1)$$

Inference is carried out by first sampling a Gaussian noise $\boldsymbol{y}_T \sim \mathcal{N}(0, \boldsymbol{I})$ and then sequentially sampling from the learned conditional distribution $p_\theta(\boldsymbol{y}_{t-1} | \boldsymbol{y}_t, \boldsymbol{x}_{tp}, \boldsymbol{x}_s)$, starting from $t = T$ and moving backward to $t = 0$.

Classifier-Free Guidance (CFG). CFG [15] can effectively shift probability mass toward data where an implicit model assigns a high likelihood to the conditioning $\boldsymbol{x}_s$ and $\boldsymbol{x}_{tp}$. It entails training the conditional and unconditional diffusion models together and combining their score estimations during inference. In practice, one randomly sets $\boldsymbol{x}_{tp} = \emptyset$ for $\eta\%$ of examples and set $\boldsymbol{x}_s = \emptyset$ for another $\eta\%$ of examples during training. During inference, two guidance scales (w_p and w_s) will be adjusted to trade off how strongly the generated samples correspond with the source image and target pose:

$$\boldsymbol{\epsilon}_{\text{cond}} = \boldsymbol{\epsilon}_{\text{uncond}} + w_p \boldsymbol{\epsilon}_{\text{pose}} + w_s \boldsymbol{\epsilon}_{\text{source}}. \quad (2)$$

Specifically, $\epsilon_{\text{uncond}} = \epsilon_\theta(\boldsymbol{y}_t, t, \emptyset, \emptyset)$ is an unconditioned prediction of the model. The pose-guided prediction ϵ_{pose} and the source-guided prediction ϵ_{source} may present different formats given different sampling strategies.

3.2. Pose-constrained latent diffusion

Model architecture. Fig. 2 illustrates the architecture of the proposed PoCoLD. Unlike PIDM [2] which directly operates in the high-dimensional pixel space, our PoCoLD learns and carries out denoising in a latent space. It consists of a pre-trained perceptual compression model with an encoder $\mathcal{E}$ and decoder $\mathcal{D}$, a latent source image encoder $\mathcal{H}_S$ and a latent prediction module $\mathcal{H}_N$. The source image encoder $\mathcal{H}_S$ encodes the latent state of the source image $\boldsymbol{z}_s = \mathcal{E}(\boldsymbol{x}_s)$, and output a stack of multi-scale feature maps $\boldsymbol{F}_s = [\boldsymbol{f}_1, \boldsymbol{f}_2, ..., \boldsymbol{f}_m]$ from different layers. Next, $\boldsymbol{F}_s$ will be sent to the UNet-based noise prediction module $\mathcal{H}_N$. The target pose condition $\boldsymbol{x}_{tp}$ is concatenated with noisy latent at each timestep. The source image condition $\boldsymbol{x}_s$ is attended via our DensePose-constrained attention. The training objective is thus rewritten as:

$$\mathcal{L}_{\text{mse}} = \mathbb{E}_{\boldsymbol{z},\mathcal{E}(\boldsymbol{x}_s),\boldsymbol{x}_{tp},\boldsymbol{\epsilon},t} \left\| \boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta \left(\boldsymbol{z}_t, t, \boldsymbol{x}_{tp}, \mathcal{E}(\boldsymbol{x}_s) \right) \right\|^2. \quad (3)$$

Pose-constrained attention. To accurately blend the texture of the source image with the noise prediction branch, we introduce pose-constrained attention. This module is integrated in multiple layers of $\mathcal{H}_N$. This is designed to drive the direction of denoising at each timestep t, so that the source texture patterns can be preserved.

Specifically, our attention module receives the multi-scale texture features $\boldsymbol{F}_s$ from $\mathcal{H}_S$ as an input and calculates the area of interest for each query position. The keys

$\boldsymbol{K}$ and values $\boldsymbol{V}$ are derived from $\mathcal{H}_S$ while queries $\boldsymbol{Q}$ are obtained from the noise features $\boldsymbol{F}_h^l$ in layer l:

$$\boldsymbol{Q} = \phi_q^l(\boldsymbol{F}_h^l), \quad \boldsymbol{K} = \phi_k^l(\boldsymbol{F}_s), \quad \boldsymbol{V} = \phi_v^l(\boldsymbol{F}_s), \quad (4)$$

where $\boldsymbol{\phi_q^l}$, $\boldsymbol{\phi_k^l}$, $\boldsymbol{\phi_v^l}$ are layer-specific linear projection layers and $\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V} \in \mathbb{R}^{hw \times d}$.

An intuitive way is to apply the widely used vanilla cross-attention [41] as:

$$\boldsymbol{A}^l = \mathtt{S}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d}}\right), \quad \boldsymbol{F}_o^l = \boldsymbol{W}_o^l \boldsymbol{A}^l \boldsymbol{V}, \quad (5)$$

where $\mathtt{S}$ refers to $\mathtt{softmax}$ and $\boldsymbol{W}_o^l$ is a learnable matrix used to generate final attended features $\boldsymbol{F}_o^l$. A residual connection is also applied: $\boldsymbol{F}_h^{l+1} = \boldsymbol{F}_o^l + \boldsymbol{F}_h^l$. However, we find that this vanilla choice is not only expensive computationally (quadratic complexity $\mathcal{O}((hw)^2 d)$), but tends to also introduce noisy attention maps since each query attends to every key (see Fig. 5).

To address these issues, we suggest to adopt sparse attention in CPIS. Specifically, the query $\boldsymbol{Q}$ and key $\boldsymbol{K}$ are both first *down-sampled* with two sets of learnable vectors $\boldsymbol{W}_q^l, \boldsymbol{W}_k^l \in \mathbb{R}^{n \times d}$. Following the *extraction then distribution* strategy [4, 30], we reformulate the dense attention (Eq. (5)) with a sparse attention $\boldsymbol{A}^l$. This is the dot product of two down-sampled attention matrices $\boldsymbol{A}^l = (\boldsymbol{A}_q^l)^\top \boldsymbol{A}_k^l$, where $\boldsymbol{A}_q^l$ and $\boldsymbol{A}_k^l$ are defined as follows[1]:

$$\boldsymbol{A}_q^l = \mathtt{S}_i\left(\frac{\boldsymbol{W}_q^l \boldsymbol{Q}^\top}{\sqrt{d}}\right), \quad \boldsymbol{A}_k^l = \mathtt{S}_j\left(\frac{\boldsymbol{W}_k^l \boldsymbol{K}^\top}{\sqrt{d}}\right), \quad (6)$$

where $\mathtt{S}_i$ and $\mathtt{S}_j$ is a $\mathtt{softmax}$ function normalizing inputs along rows and columns, respectively. This operation reduces the computation complexity to $\mathcal{O}(hwnd + (hw)^2 n)$, which is $n/hw + n/d$ of the vanilla attention. Since $n \ll d$ and $n \ll hw$ in practice, this sparse attention can significantly reduce resource usage. Importantly, this avoids exhaustive pairing between $\boldsymbol{Q}$ and $\boldsymbol{K}$, and the attention redundancy can be alleviated (Fig. 5).

Typically, the pose deformation is *implicitly* estimated using the content of the source image $\boldsymbol{x}_s$ and target pose $\boldsymbol{x}_{tp}$ [2, 30]. This makes the synthesis less controllable. To remedy this, we exploit the structural body part information (*e.g.*, I map) of DensePose. We derive a novel attention constraint with the relationship between the target pose $\boldsymbol{x}_{tp}$ and source pose $\boldsymbol{x}_{sp}$. More specifically, given flattened $\overline{\boldsymbol{x}_{tp}}, \overline{\boldsymbol{x}_{sp}} \in \mathbb{R}^{hw}$ at one layer l, we first get a binary constraint map $\boldsymbol{C}^l \in \mathbb{R}^{hw \times hw}$, which computes the element-wise equality between the body part labels of two poses:

$$\boldsymbol{C}_{i,j}^l = [\overline{\boldsymbol{x}_{tp}}_i = \overline{\boldsymbol{x}_{sp}}_j]. \quad (7)$$

[1]We omit multi-head processes for a clear demonstration.

This can be considered as a coarse semantic attention mask. An intuition is that two pixels with the same body part label should be attended to each other. To that end, we design a regularization term to minimize the attention scores of unmatched body parts, which is formulated as:

$$\mathcal{L}_{\text{p}}^l = \left\|\boldsymbol{A}_{i,j}^l\right\|^2 \text{ s.t. } \boldsymbol{C}_{i,j}^l = 0 \text{ and } \sum_j \boldsymbol{C}_i^l \neq 0, \quad (8)$$

where we down-sample $\boldsymbol{C}^l$ with the nearest-neighbor interpolation to meet the same resolution of $\boldsymbol{A}^l$ in each layer. Finally, the overall learning objective of our model is the noise prediction loss and the weighted summation of regularization terms from all layers: $\mathcal{L} = \mathcal{L}_{\text{mse}} + \lambda \sum \mathcal{L}_{\text{p}}^l$.

Sampling techniques. To amplify the conditional signal of $\boldsymbol{x}_s$ and $\boldsymbol{x}_{tp}$ in the sampled images, we adopt CFG in our training and inference stage. In case of $\boldsymbol{x}_s = \emptyset$, instead of replacing $\boldsymbol{x}_s$ with all zeros [2], we let the noise prediction module $\mathcal{H}_N$ *skip* the attention mechanism to speed up denoising process. During training, $\mathcal{H}_N$ thus degrades to a single-conditioned model in this case, *i.e.*, Eq. (3) is replaced with $\mathcal{L}_{\text{mse}} = \mathbb{E}_{\boldsymbol{z},\boldsymbol{x}_p,\boldsymbol{\epsilon},t} \left\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta\left(\boldsymbol{z}_t, t, \boldsymbol{x}_p\right)\right\|^2$. Additionally, instead of disentangled CFG [2], we use cumulative CFG [3] for Eq. (2) during inference, where $\epsilon_{\text{pose}} = \epsilon_\theta(\boldsymbol{y}_t, t, \boldsymbol{x}_{tp}, \emptyset) - \epsilon_{\text{uncond}}$ and $\epsilon_{\text{source}} = \epsilon_\theta(\boldsymbol{y}_t, t, \boldsymbol{x}_{tp}, \boldsymbol{x}_s) - \epsilon_\theta(\boldsymbol{y}_t, t, \boldsymbol{x}_{tp}, \emptyset)$.

4. Experiments

Datasets. We carry out experiments on the DeepFashion In-shop Clothes Retrieval Benchmark [20]. This dataset includes 52,712 high-resolution person images in the fashion domain. Following the same data configuration as [51], we split DeepFashion into non-overlapped training and test subsets with 101,966 and 8,570 pairs, respectively. Each pair includes the same person in the same garments but with different poses/viewpoints. The DensePose map of each image is registered with the off-the-shelf model [10].

Objective metrics. We evaluate the performance of the model using three different metrics, following previous works. The *Structure Similarity Index Measure* (SSIM) [44] and the *Learned Perceptual Image Patch Similarity* (LPIPS) [48] are utilized to evaluate the accuracy of reconstruction. SSIM determines the similarity of images at the pixel-level, while LPIPS measures the perceptual distance by leveraging a network trained on human judgments. Additionally, we use the *Frèchet Inception Distance* (FID) [13] to quantify the realism of the generated images. FID calculates the Wasserstein-2 distance between the distributions of the synthesized images and those of the real images.

Subjective metric. The objective metrics (especially FID) are known to sometimes have a weak correlation with the actual generated image quality, which ultimately needs to be judged by humans. We therefore conducted a user study

involving 15 human participants. We presented the volunteers with 30 pairs of randomly selected and shuffled images, including the source image, target pose, ground truth, and images generated by our method and seven baselines [2,25,30,31,46,47,49]. The participants were asked to choose the image that appeared most realistic and plausible with respect to the source image and target pose. We quantified the results of this study using a metric called *Jab* [2,49], which represents the percentage of images that were considered the best among all generations.

Implementation details. We use the off-the-shelf ft-MSE autoencoder of Stable Diffusion as our latent autoencoder. Our PoCoLD is implemented with PyTorch [28] and HuggingFace Diffusers [42], which has been trained with $T = 1,000$ denoising steps and a linear noise schedule. In all experiments, we use a total batch size of 32 on 2 Tesla V100 GPUs with `fp16` precision. Adam optimizer [17] is used with the learning rate set to 5×10^{-5} and the linear warm-up from zero at the first 1,000 steps. For inference sampling, PNDM scheduler [19] is used with 50 steps and the η, w_p, w_s of our CFG is set to 10, 5, 5, respectively. More implementation details are given in the supplementary file.

4.1. Quantitative and qualitative comparisons

Comparative results. We quantitatively compare objective and subjective metrics between our proposed PoCoLD and representative prior arts, including Def-GAN [35], PATN [51], ADGAN [25], PISE [46], GFLA [31], DPTN [47], CocosNet2 [50], NTED [30] and PIDM [2]. We also calculate objective metrics for the ground truths and auto-encoder reconstructions to have a clear reference of our current position. The evaluations are done on both 256×176 and 512×352 resolutions for DeepFashion [20].

From the results in Tab. 1, the following observations can be made. (1) Our PoCoLD outperforms other methods in terms of reconstruction metrics such as SSIM and LPIPS. This demonstrates that our model is capable of producing images that not only exhibit accurate structures but also correctly transfer the texture of the source image to the target pose. (2) It is noted that on FID, our PoCoLD is clearly inferior to that of PIDM, despite beating all other competitors. However, a close inspection suggests that this is largely due to the fact that PIDM might have overfitted to the training data. More specifically, there seems to be a quite big distribution shift between the ground truths test set and the training set – the FID score of the ground truth for the test set is approximately 8 which is very close to ours, but also much worse than that of PIDM. We thus conclude that the learned distribution of PIDM is overfitted towards the training set, and FID score in this case is not the best reflection of model performance. (3) Our model's superiority is corroborated by the highest *Jab* score, which indicates that our method aligns well with human perception.

Resolution	Methods	FID↓	SSIM↑	LPIPS↓	Jab↑
256×176	Def-GAN [35]	18.457	0.6786	0.2330	-
	PATN [51]	20.751	0.6709	0.2562	-
	ADGAN [25]	14.458	0.6721	0.2283	1.56
	PISE [46]	13.610	0.6629	0.2059	2.89
	GFLA [31]	10.573	0.7074	0.2341	7.56
	DPTN [47]	11.387	0.7112	0.1931	1.11
	CASD [49]	11.373	<u>0.7248</u>	0.1936	5.78
	NTED [30]	8.6838	0.7182	0.1752	15.33
	PIDM* [2]	**6.4362**	0.7109	<u>0.1685</u>	<u>30.67</u>
	PoCoLD (Ours)	<u>8.0667</u>	**0.7310**	**0.1642**	**35.11**
	AE Recon.	8.2335	0.9668	0.0110	-
	Ground Truth	7.8700	1.0000	0.0000	-
512×352	CocosNet2 [50]	13.325	0.7236	0.2265	-
	NTED [30]	**7.7821**	<u>0.7376</u>	<u>0.1980</u>	-
	PoCoLD (Ours)	<u>8.4163</u>	**0.7430**	**0.1920**	-
	AE Recon.	8.5750	0.9230	0.0260	-
	Ground Truth	8.0198	1.0000	0.0000	-

Table 1. Quantitative comparison of the proposed PoCoLD with several state-of-the-art models in terms of *Frèchet Inception Distance* (FID), *Structure Similarity Index Measure* (SSIM), *Learned Perceptual Image Patch Similarity* (LPIPS) and users' feedback (Jab). The results are shown on both 256×176 and 512×352 resolutions for DeepFashion [20]. The best/second results are shown in bold/underlined. PIDM* is evaluated on the generated images released by the authors.

Methods	Model Size↓	Inference Speed↓	Memory Use↓
PIDM [2]	688.00MB	9.25±0.54s	6639MB
PoCoLD (Ours)	**395.89MB**	**4.99±0.02s**	**1855MB**

Table 2. Efficiency comparison between our PoCoLD and prior art diffusion-based model (PIDM [2]). The results are obtained using a machine with one single Tesla V100 GPU. The inference speed is measured over 10 times for one 256×176 image generation with 50 denoising steps.

Efficiency comparisons. Table 2 presents a comparison of three key efficiency metrics (model size, inference speed, and memory use) between our proposed PoCoLD and PIDM [2] under the same number of denoising steps. The denoising steps of our PoCoLD are conducted in the latent space, and the noise prediction module receives compressed latent images as input during inference. In contrast, PIDM operates on the pixel space at every time step, resulting in slower generation speed and higher GPU memory usage. Statistically, our PoCoLD achieves doubled generation speed and 3.6 times lower memory use while maintaining a model size of only 57.5% of PIDM's size, demonstrating our superior computation efficiency.

Qualitative comparisons. We provide a comprehensive visual comparison of our method with other state-of-the-art techniques on the DeepFashion dataset in Fig. 3. The results of others are obtained using pre-trained models or released

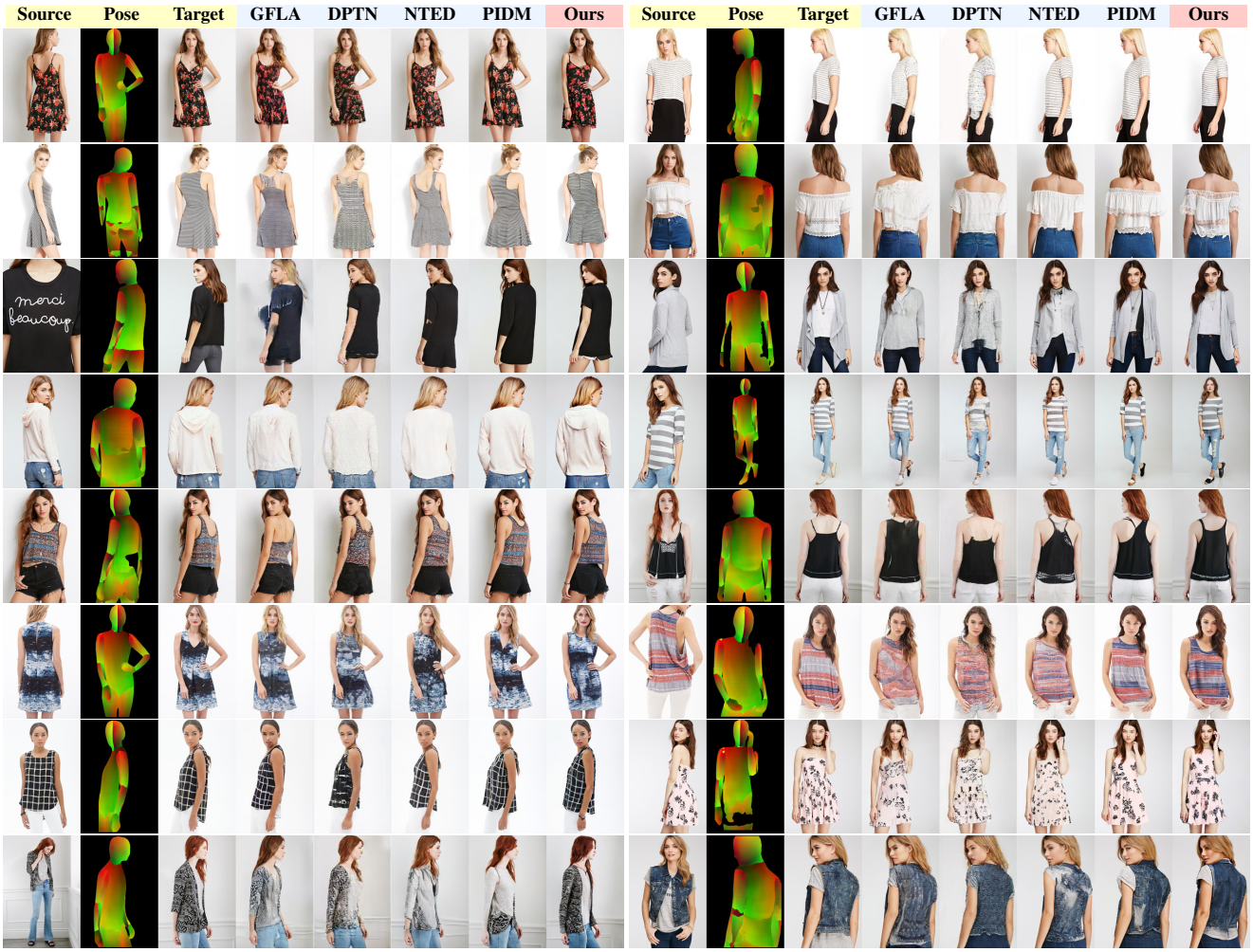

Figure 3. Qualitative comparisons with several state-of-the-art models on the DeepFashion dataset [20]. The model input includes both the target pose $\boldsymbol{x}_p$ and the source image $\boldsymbol{x}_s$. From left to right are the results by GFLA [31], DPTN [47], NTED [30], PIDM [2], and ours respectively. (Best viewed when zoom in)

images provided by the corresponding authors. Our visual comparison in Fig.3 reveals several key observations: (1) Although GFLA [31] adopts appearance flow to preserve the texture in the source image, it struggles to obtain reasonable results for the invisible regions of the source image. (2) DPTN [47] and NTED [30] slightly improve results by using attention mechanisms, but they still fall short in complex scenarios, as seen in the last three rows. (3) Compared with the aforementioned GAN-based methods, diffusion-based approaches like PIDM [2] and our PoCoLD excel at retaining the source appearance while producing more natural and sharper images. (4) In comparison to PIDM, our PoCoLD predicts more accurate pose deformation, as evidenced by the right column. This is because we utilize DensePose as our condition and use it to constrain our learned attention.

4.2. Ablation study

We perform several ablation studies to validate the merits of the proposed pose-constrained attention. Specifically, we design two baselines to compare with. `B1` is a latent diffusion model using vanilla cross-attention [41] to extracting source image textures, while `B2` is using our down-sampled attention but without pose constraints. We train all ablation models with the same setting as that of our model.

The quantitative results of the ablation study are shown in Tab. 3. With the help of the diffusion model, `B1` achieves satisfactory results and outperforms several GAN-based methods, as demonstrated in Tab. 1. Our down-sampled attention, implemented in place of the vanilla cross-attention, led to noteworthy enhancements in performance metrics, as evidenced by `B2`. This highlights the efficacy of down-

Methods	Memory Use↓	Speed↓	FID↓	SSIM↑	LPIPS↓
B1	4969MB	5.30±0.06s	8.2903	0.7095	0.1783
B2	**1855MB**	**4.99±0.02s**	**8.0285**	0.7241	0.1701
Ours	**1855MB**	**4.99±0.02s**	8.0667	**0.7310**	**0.1642**

Table 3. Ablation on the impact of pose-constrained attention (B1: vanilla cross-attention, B2: down-sampled attention w/o pose-constraints). All models are trained and evaluated under identical settings. The results are obtained using 256×176 images.

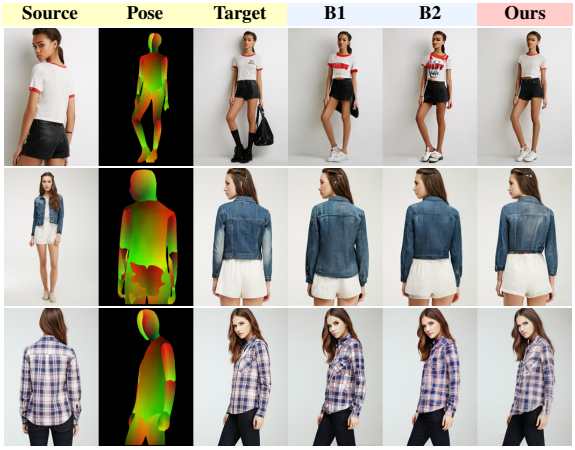

Figure 4. Qualitaive results of ablation study. The images correspond to the ablation studies in Tab. 3. Our method is superior in detail, shape, and texture preservation over the baselines.

sampled sparse attention. Additionally, B2 offers a memory usage reduction of 63% and improves inference speed by 6%. Moreover, our pose constraints further enhance the model's results on the reconstruction metrics, proving the effectiveness of our pose constraints in accurately predicting pose deformation.

We also show the qualitative results in Fig. 4. The results show that B1 struggles to maintain the shape of the garments. This is primarily due to the vanilla cross-attention mechanism that allows each query to attend to every key, leading to imprecise localization of the condition. Although B2 improves shape preservation, it still falls short in terms of accurately preserving the texture and details due to the lack of pose constraints, while our model can faithfully reconstruct the textures of source images.

To have more intuitive understanding of what attention is learned, we visualize the attention maps at the mid-layer for both baselines and our model. As shown in Fig. 5, our model enables the query more accurately attend to the related regions in the source image than both baselines. For example, our attention isolates the hat region (second row left) and the right arm (third row left) more reliably.

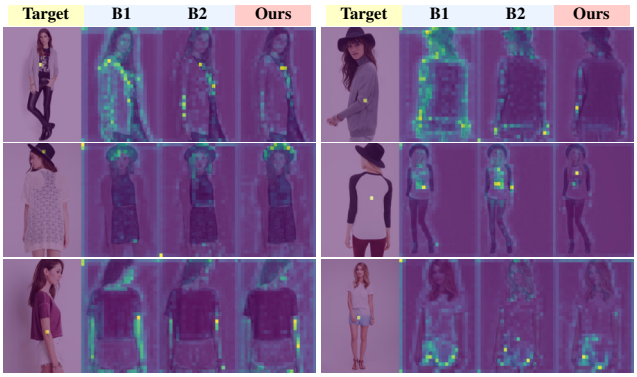

Figure 5. Visualizing the attention from the mid-layer of noise prediction module, captured at the resolution of 32. Attention maps are averaged over all heads and 50 time steps. The first image in each group stands for the query in the generation branch, while the rest images are the key from corresponding source images.

Figure 6. Qualitative results of PoCoLD for pose-only conditioned image generation using different random seeds. In this setting, the noise prediction module skips the attention process. This demonstrates that our model can generate diverse images.

4.3. More applications

Unlike existing works [30] that typically require different models for various tasks, such as pose-only generation and appearance control, our model can perform multiple tasks with a single unified model.

Pose-only conditioned person image generation. When the source image $\boldsymbol{x}_s$ is not given, controllable person image synthesis degrades to a generation problem with only pose condition $\boldsymbol{x}_p$. With a higher degree of freedom, this setting can generate more diverse person images with different appearances. Since we are setting $\boldsymbol{x}_s = \emptyset$ with a certain probability for classifier-free guidance, it is natural for our PoCoLD to do pose-only conditioning generation. As mentioned in Sec. 3.1, we let the noise prediction module $\mathcal{H}_N$ skip the attention process and only condition on $\boldsymbol{x}_p$. During

Figure 7. Qualitative results of PoCoLD for appearance transfer. The garment's appearance in the reference image can be controlled while maintaining the person's pose and identity. This is achieved by masking out regions of interest in the reference images while keeping other parts intact during inference, without the need for additional training.

sampling, a standard classifier-free guidance [15] is applied: $\boldsymbol{\epsilon}_{\text{cond}} = \boldsymbol{\epsilon}_{\text{uncond}} + w_p \boldsymbol{\epsilon}_{\text{pose}}$. We show some pose-only conditioned generation results in Fig. 6, where the diverse and reasonable appearances are observed.

Appearance transfer. Our PoCoLD can also achieve explicit appearance control by spatially interpolating the appearances of different references. The task is to generate an image $\bar{\boldsymbol{y}}$ that is consistent with the source (style) image $\boldsymbol{x}_s$ in the masked region of interest marked by a binary mask $\boldsymbol{m}$ in the reference image $\boldsymbol{y}^{ref}$, while keeping the rest of the image unchanged. This can be achieved using our trained diffusion model that predicts $\boldsymbol{y}_t$ iteratively from a Gaussian noise $\boldsymbol{y}_T \sim \mathcal{N}(0, \boldsymbol{I})$ in each step t. Following PIDM [2], the binary mask $\boldsymbol{m}$ is used to retain the unmasked regions of $\boldsymbol{y}^{ref}$, and the relation $\boldsymbol{y}_t = \boldsymbol{m} \odot \boldsymbol{y}_t + (1 - \boldsymbol{m}) \odot \boldsymbol{y}_t^{ref}$ is applied at each step t, where $\boldsymbol{y}_t^{ref}$ is the noisy version of $\boldsymbol{y}^{ref}$ at step t. Our method enables appearance transfer without using any additional human parser maps, and the resulting output images exhibit coherent textures and seamlessly combined areas of interest, as demonstrated in Fig. 7.

Figure 8. Typical failure cases caused by exaggerated pose (top left), underrepresented garments (top right), noisy pose annotation (bottom left), and garment change (bottom right).

4.4. Failure cases and limitations

Failure cases. Despite achieving satisfactory results in most cases, our model sometimes produces unsatisfactory outcomes as demonstrated in Fig. 8, where inconsistencies or artifacts are visible. We have identified four specific scenarios that invariably lead to failure in generating realistic outputs: **(1)** poses that are exaggerated and vastly different from the ones in the training set; **(2)** garments that are underrepresented and vary significantly across viewpoints; **(3)** noisy DensePose estimations; and **(4)** changes in the garment between the source and target images. To address this issue, we believe that employing a more sophisticated DensePose estimator and training on a more diverse dataset would mitigate these problems.

Limitations. **(1)** To compress the image into latent space, we used a pre-trained auto-encoder [32] that down-scales the image resolution by 8 times, leading to the loss of information, as quantitatively shown in Tab. 1. This affects the reconstruction of important appearances such as face identity and garment texture, especially when the region of interest is small compared to the whole image. We suggest that using auto-encoders with smaller down-sample ratios may help alleviate this problem. **(2)** In addition, our method does not enforce multi-view-consistent generation (*e.g.*, consistent video generation across different frames) due to the lack of multi-view supervision during training.

5. Conclusions

We have presented PoCoLD for controllable person image synthesis in a latent diffusion paradigm. We introduce an efficient *pose-constrained attention* to alleviate attention redundancy and explicitly model the pose deformation between source and target images. Our model achieves superior performance and generates high-resolution realistic images with much less memory usage and quicker inference speed than the previous diffusion-based method. Meanwhile, PoCoLD also enables further applications like pose-only conditioned generation and appearance transfer.

References

[1] Badour AlBahar, Jingwan Lu, Jimei Yang, Zhixin Shu, Eli Shechtman, and Jia-Bin Huang. Pose with style: Detail-preserving pose-guided image synthesis with conditional stylegan. *ACM TOG*, 2021. 2

[2] Ankan Kumar Bhunia, Salman Khan, Hisham Cholakkal, Rao Muhammad Anwer, Jorma Laaksonen, Mubarak Shah, and Fahad Shahbaz Khan. Person image synthesis via denoising diffusion model. In *CVPR*, 2023. 1, 2, 3, 4, 5, 6, 8

[3] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*, 2023. 2, 4

[4] Yunpeng Chen, Yannis Kalantidis, Jianshu Li, Shuicheng Yan, and Jiashi Feng. Aˆ 2-nets: Double attention networks. In *NeurIPS*, 2018. 4

[5] Aiyu Cui, Daniel McKee, and Svetlana Lazebnik. Dressing in order: Recurrent person image generation for pose transfer, virtual try-on and outfit editing. In *ICCV*, 2021. 2

[6] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *NeurIPS*, 2020. 1, 2, 3

[7] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, 2021. 2

[8] Patrick Esser, Ekaterina Sutter, and Björn Ommer. A variational u-net for conditional appearance and shape generation. In *CVPR*, 2018. 2

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 2020. 1, 2

[10] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *CVPR*, 2018. 2, 4

[11] Xiao Han, Licheng Yu, Xiatian Zhu, Li Zhang, Yi-Zhe Song, and Tao Xiang. Fashionvil: Fashion-focused vision-and-language representation learning. In *ECCV*, 2022. 1

[12] Xiao Han, Xiatian Zhu, Licheng Yu, Li Zhang, Yi-Zhe Song, and Tao Xiang. Fame-vil: Multi-tasking vision-language model for heterogeneous fashion tasks. In *CVPR*, 2023. 1

[13] 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. In *NeurIPS*, 2017. 4

[14] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 1, 2, 3

[15] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS Workshop*, 2021. 2, 3, 8

[16] Diederik Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. In *NeurIPS*, 2021. 2

[17] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 5

[18] Yining Li, Chen Huang, and Chen Change Loy. Dense intrinsic appearance flow for human pose transfer. In *CVPR*, 2019. 1, 2

[19] Luping Liu, Yi Ren, Zhijie Lin, and Zhou Zhao. Pseudo numerical methods for diffusion models on manifolds. In *ICLR*, 2022. 5

[20] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *CVPR*, 2016. 2, 4, 5, 6

[21] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM TOG*, 2015. 2

[22] Zhengyao Lv, Xiaoming Li, Xin Li, Fu Li, Tianwei Lin, Dongliang He, and Wangmeng Zuo. Learning semantic person image generation by region-adaptive normalization. In *CVPR*, 2021. 2

[23] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. Pose guided person image generation. In *NeurIPS*, 2017. 1, 2

[24] Liqian Ma, Qianru Sun, Stamatios Georgoulis, Luc Van Gool, Bernt Schiele, and Mario Fritz. Disentangled person image generation. In *CVPR*, 2018. 2

[25] Yifang Men, Yiming Mao, Yuning Jiang, Wei-Ying Ma, and Zhouhui Lian. Controllable person image synthesis with attribute-decomposed gan. In *CVPR*, 2020. 1, 2, 5

[26] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. 1, 2

[27] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *ICML*, 2022. 2

[28] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019. 5

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

[30] Yurui Ren, Xiaoqing Fan, Ge Li, Shan Liu, and Thomas H Li. Neural texture extraction and distribution for controllable person image synthesis. In *CVPR*, 2022. 1, 2, 4, 5, 6, 7

[31] Yurui Ren, Xiaoming Yu, Junming Chen, Thomas H Li, and Ge Li. Deep image spatial transformation for person image generation. In *CVPR*, 2020. 1, 2, 5, 6

[32] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 1, 2, 3, 8

[33] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022. 1, 2

[34] Kripasindhu Sarkar, Vladislav Golyanik, Lingjie Liu, and Christian Theobalt. Style and pose control for image synthesis of humans from a single monocular view. *arXiv preprint arXiv:2102.11263*, 2021. 2

[35] Aliaksandr Siarohin, Enver Sangineto, Stéphane Lathuiliere, and Nicu Sebe. Deformable gans for pose-based human image generation. In *CVPR*, 2018. 1, 2, 5

[36] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, 2015. 1, 2, 3

[37] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021. 2

[38] Hao Tang, Song Bai, Li Zhang, Philip HS Torr, and Nicu Sebe. Xinggan for person image generation. In *ECCV*, 2020. 2

[39] Jilin Tang, Yi Yuan, Tianjia Shao, Yong Liu, Mengmeng Wang, and Kun Zhou. Structure-aware person image generation with pose decomposition and semantic correlation. In *AAAI*, 2021. 2

[40] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. In *NeurIPS*, 2017. 2

[41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 3, 4, 6

[42] Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers, 2022. 5

[43] Andrey Voynov, Kfir Aberman, and Daniel Cohen-Or. Sketch-guided text-to-image diffusion models. *arXiv preprint arXiv:2211.13752*, 2022. 2

[44] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE TIP*, 2004. 4

[45] 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 *CVPR*, 2023. 2

[46] Jinsong Zhang, Kun Li, Yu-Kun Lai, and Jingyu Yang. Pise: Person image synthesis and editing with decoupled gan. In *CVPR*, 2021. 2, 5

[47] Pengze Zhang, Lingxiao Yang, Jian-Huang Lai, and Xiaohua Xie. Exploring dual-task correlation for pose guided person image generation. In *CVPR*, 2022. 2, 5, 6

[48] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 4

[49] Xinyue Zhou, Mingyu Yin, Xinyuan Chen, Li Sun, Changxin Gao, and Qingli Li. Cross attention based style distribution for controllable person image synthesis. In *ECCV*, 2022. 2, 5

[50] Xingran Zhou, Bo Zhang, Ting Zhang, Pan Zhang, Jianmin Bao, Dong Chen, Zhongfei Zhang, and Fang Wen. Cocosnet v2: Full-resolution correspondence learning for image translation. In *CVPR*, 2021. 2, 5

[51] Zhen Zhu, Tengteng Huang, Baoguang Shi, Miao Yu, Bofei Wang, and Xiang Bai. Progressive pose attention transfer for person image generation. In *CVPR*, 2019. 4, 5