

## Person Image Synthesis via Denoising Diffusion Model

Ankan Kumar Bhunia<sup>1</sup> Salman Khan<sup>1,2</sup> Hisham Cholakkal<sup>1</sup> Rao Muhammad Anwer<sup>1,4</sup>  
 Jorma Laaksonen<sup>4</sup> Mubarak Shah<sup>5</sup> Fahad Shahbaz Khan<sup>1,3</sup>

<sup>1</sup>Mohamed bin Zayed University of AI, UAE <sup>2</sup>Australian National University, Australia  
<sup>3</sup>Linköping University, Sweden <sup>4</sup>Aalto University, Finland <sup>5</sup>University of Central Florida, USA

### Abstract

The pose-guided person image generation task requires synthesizing photorealistic images of humans in arbitrary poses. The existing approaches use generative adversarial networks that do not necessarily maintain realistic textures or need dense correspondences that struggle to handle complex deformations and severe occlusions. In this work, we show how denoising diffusion models can be applied for high-fidelity person image synthesis with strong sample diversity and enhanced mode coverage of the learnt data distribution. Our proposed Person Image Diffusion Model (PIDM) disintegrates the complex transfer problem into a series of simpler forward-backward denoising steps. This helps in learning plausible source-to-target transformation trajectories that result in faithful textures and undistorted appearance details. We introduce a ‘texture diffusion module’ based on cross-attention to accurately model the correspondences between appearance and pose information available in source and target images. Further, we propose ‘disentangled classifier-free guidance’ to ensure close resemblance between the conditional inputs and the synthesized output in terms of both pose and appearance information. Our extensive results on two large-scale benchmarks and a user study demonstrate the photorealism of our proposed approach under challenging scenarios. We also show how our generated images can help in downstream tasks. Code is available at <https://github.com/ankanbhunia/PIDM>.

### 1. Introduction

The Pose-guided person image synthesis task [19, 23, 30] aims to render a person’s image with a desired pose and appearance. Specifically, the appearance is defined by a given source image and the pose by a set of keypoints. Having control over the synthesized person images in terms of pose and style is an important requisite for applications such as e-commerce, virtual reality, metaverse and content generation for the entertainment industry. Furthermore, the generated



Figure 1. (a) Our proposed PIDM is a denoising diffusion model where the generative path is conditioned on the pose and style. PIDM breaks down the problem into a series of forward-backward diffusion steps to learn the plausible transfer trajectories. (b) Comparison of PIDM with the recently introduced NTED [18]. PIDM accurately retains the appearance of the source style image while also producing images that are more natural and sharper while NTED struggles to adequately preserve the source appearance in complex scenarios (marked in red boxes).

images can be used to improve performance on downstream tasks such as person re-identification [30]. The challenge is to generate photorealistic outputs tightly conforming with the given pose and appearance information.

In the literature, person synthesis problem is generally tackled using Generative Adversarial Networks (GAN) [4] which try to generate a person in a desired pose using a single forward pass. However, preserving a coherent structural, appearance and global body composition in the new pose is a challenging task to achieve in one shot. The resulting outputs commonly experience deformed textures and unrealistic body shapes, especially when synthesizing occluded body parts (see Fig. 1). Further, GANs are prone to unstable training behaviour due to adversarial min-max objective and lead to limited diversity in the generated samples. Similarly, Variational Autoencoder [8] based solutions

have been explored that are relatively stable, but suffer from blurry details and offer low-quality outputs than GANs due to their dependence on a surrogate loss for optimization.

In this work, we frame the person synthesis problem as a series of diffusion steps that progressively transfer a person in the source image to the target pose. Diffusion models [6] are motivated from non-equilibrium thermodynamics that define a Markov chain of slowly adding noise to the input samples (forward pass) and then reconstructing the desired samples from noise (reverse pass). In this manner, rather than modeling the complex transfer characteristics in a single go, our proposed person synthesis approach PIDM breaks down the problem into a series of forward-backward diffusion steps to learn the plausible transfer trajectories. Our approach can model the intricate interplay of the person’s pose and appearance, offers higher diversity and leads to photorealistic results without texture deformations (see Fig. 1). In contrast to existing approaches that deal with major pose shifts by requiring parser maps denoting human body parts [14, 23, 28], or dense 3D correspondences [9, 10] to fit human body topology by warping the textures, our approach can learn to generate realistic and authentic images without such detailed annotations.

Our major contributions are as follows:

- We develop the first *diffusion-based approach* for pose-guided person synthesis task which can work under challenging pose transformations while preserving appearance, texture and global shape characteristics.
- To effectively model the complex interplay between appearance and pose information, we propose a *texture diffusion module*. This module exploits the correspondences between source and target appearance and pose details, hereby obtaining artefact free images.
- In the sampling procedure, we introduce *disentangled classifier-free guidance* to tightly align the output image style and pose with the source image appearance and target pose, respectively. It ensures close resemblance between the conditions that are input to the generative model and the generated output.
- Our results on DeepFashion [11] and Market-1501 [27] benchmarks set new state of the art. We also report a user study to evaluate the qualitative features of generated images. Finally, we demonstrate that synthesized images can be used to improve performance in downstream tasks *e.g.*, person re-identification.

## 2. Related Work

**Pose-guided Person Image Synthesis:** The problem of human pose transfer has been studied extensively during the recent years, especially with the unprecedented success of GAN-based models [15] for conditional image synthesis. An early attempt [13] proposes a coarse-to-fine approach

to first generate a rough image with the target pose and then refine the results adversarially. The method simply concatenates the source image, the source pose, and the target pose as inputs to obtain the target image, which leads to feature misalignment. To address this issue, Essner *et al.* [3] attempt to disentangle the appearance and the pose of person images using VAE-based design and a UNet based skip-connection architecture. Siarohin *et al.* [20] improve the model by introducing deformable skip connections to spatially transform the textures, which decomposes the overall deformation by a set of local affine transformations. Subsequently, some works [9, 10, 19] use flow-based deformation to transform the source information to improve pose alignment. Ren *et al.* [19] propose GFLA that obtains the global flow fields and occlusion mask, which are used to warp local patches of the source image to match the required pose. Another group of works [9, 10] use geometric models that fit a 3D mesh human model onto the 2D image, and subsequently predict the 3D flow, which finally warps the source appearance. On the other hand, without any deformation operation, Zhu *et al.* [30] propose to progressively transform the source image by a sequence of transfer blocks. However, useful information can be lost during multiple transfers, which may result in blurry details. ADGAN [14] uses a texture encoder to extract style vectors for human body parts and gives them to several AdaIN residual blocks to synthesize the final image. Methods such as PISE [23], SPGNet [12] and CASD [28] make use of parsing maps to generate the final image. CoCosNet [25, 29] extracts dense correspondences between cross-domain images with attention-based operation. Recently, Ren *et al.* [18] propose a framework NTED based on neural texture extraction and distribution operation, which achieves superior results.

**Diffusion Models:** The existing GAN-based approaches attempt to directly transfer the style of the source image to a given target pose, which requires the architecture to model complex transformation of pose. In this work, we present a diffusion-based framework named PIDM that breaks the pose transformation process into several conditional denoising diffusion steps, in which each step is relatively simple to model. Diffusion models [6] are recently proposed generative models that can synthesize high-quality images. After success in unconditional generation, these models are extended to work in conditional generation settings, demonstrating competitive or even better performance than GANs. For class-conditioned generation, Dhariwal *et al.* [2] introduce classifier-guided diffusion, which is later adapted by GLIDE [16] to enable conditioning over CLIP textual representations. Recently, Ho *et al.* [7] propose a Classifier-Free Guidance approach that enables conditioning without requiring pretraining of the classifiers. In this work, we develop the first *diffusion-based* approach for pose-guided person synthesis task. We also introduce *disentan-*

gled classifier-free guidance to tightly align the output image style and pose with the source image appearance and target pose, respectively.

### 3. Proposed Method

**Motivation:** The existing pose-guided person synthesis methods [12, 14, 18, 19, 23, 25, 28] rely on GAN-based frameworks where the model attempts to directly transfer the style of the source image into a given target pose in a single forward pass. It is quite challenging to directly capture the complex structure of the spatial transformation, therefore current CNN-based architectures often struggle to transfer the intricate details of the cloth texture patterns. As a result, the existing methods yield noticeable artifacts, which become more evident when the generator needs to infer occluded body regions from the given source image.

Motivated by these observations, we advocate that instead of learning the complex structure directly in a single step, deriving the final image using successive intermediate transfer steps can make the learning task simpler. To enable the above progressive transformation scheme, we introduce *Person Image Diffusion Model* or PIDM, a diffusion-based [6] person image synthesis framework that breaks down the generation process into several conditional denoising diffusion steps, each step being relatively simple to model. A single step in the diffusion process can be approximated by a simple isotropic Gaussian distribution. We observe that our diffusion-based texture transfer technique PIDM can bring the following benefits: **(1) High-quality synthesis:** Compared to previous GAN-based approaches, PIDM generates photo-realistic results when dealing with complex cloth-textures and extreme pose angles. **(2) Stable training:** Existing GAN-based approaches use multiple loss objectives alongside adversarial losses, which are often difficult to balance, resulting in unstable training. In contrast, PIDM exhibits better training stability and mode coverage. Also, our model is less prone to hyperparameters. **(3) Meaningful interpolation:** Our proposed PIDM allows us to achieve smooth and consistent linear interpolation in the latent space. **(4) Flexibility:** The models from existing work are usually task dependent, requiring different models for various tasks (e.g., separate models for unconditional, pose-conditional, pose and style-conditional generation tasks). In contrast, in our case, a single model can be used to perform multiple tasks. Furthermore, PIDM inherits the flexibility and controllability of diffusion models that enable various downstream tasks (e.g., appearance control, see Fig. 7) using our model.

**Overall Framework:** Fig. 2 shows the overview of the proposed generative model. Given a source image  $\mathbf{x}_s$  and a target pose  $\mathbf{x}_p$ , our goal is to train a conditional diffusion model  $p_\theta(\mathbf{y}|\mathbf{x}_s, \mathbf{x}_p)$  where the final output image  $\mathbf{y}$  should not only satisfy the target pose matching requirement, but

should also have the same style as in  $\mathbf{x}_s$ .

The denoising network  $\epsilon_\theta$  in PIDM is a UNet-based design composed of a noise prediction module  $\mathcal{H}_N$  and a texture encoder  $\mathcal{H}_E$ . The encoder  $\mathcal{H}_E$  encodes the texture patterns of the source image  $\mathbf{x}_s$ . To obtain multi-scale features we derive output from the different layers of  $\mathcal{H}_E$  resulting in stacked feature representation  $\mathbf{F}_s = [\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_m]$ . To transfer rich multi-scale texture patterns from the source image distribution to the noise prediction module  $\mathcal{H}_N$ , we propose to use cross-attention based *Texture diffusion blocks* (TDB) that are embedded in different layers of  $\mathcal{H}_N$ . This allows the network to fully exploit the correspondences between the source and target appearances, thus resulting in distortion-free images. During inference, to amplify the conditional signal of  $\mathbf{x}_s$  and  $\mathbf{x}_p$  in the sampled images, we adapt the classifier-free guidance [7] in our sampling technique to achieve *disentangled guidance*. It not only improves the overall quality of the generation, but also ensures accurate transfer of texture patterns. We provide detailed analysis of the proposed generative model in Sec. 3.1, the Texture diffusion blocks in Sec. 3.2 and our disentangled guidance based sampling technique in Sec. 3.3.

#### 3.1. Texture-Conditioned Diffusion Model

The generative modeling scheme of PIDM is based on the Denoising diffusion probabilistic model [6] (DDPM). The general idea of DDPM is to design a diffusion process that gradually adds noise to the data sampled from the target distribution  $\mathbf{y}_0 \sim q(\mathbf{y}_0)$ , while the backward denoising process attempts to learn the reverse mapping. The denoising diffusion process eventually converts an isotropic Gaussian noise  $\mathbf{y}_T \sim \mathcal{N}(0, \mathbf{I})$  into the target data distribution in  $T$  steps. Essentially, this scheme divides a complex distribution-modeling problem into a set of simple denoising problems. The forward diffusion path of DDPM is a Markov chain with the following conditional distribution:

$$q(\mathbf{y}_t | \mathbf{y}_{t-1}) = \mathcal{N}(\mathbf{y}_t; \sqrt{1 - \beta_t} \mathbf{y}_{t-1}, \beta_t \mathbf{I}). \quad (1)$$

where  $t \sim [1, T]$  and  $\beta_1, \beta_2, \dots, \beta_T$  is a fixed variance schedule with  $\beta_t \in (0, 1)$ . Using the notation  $\alpha_t = 1 - \beta_t$  and  $\bar{\alpha}_t = \prod_{i=1}^t \alpha_i$ , we can sample from  $q(\mathbf{y}_t | \mathbf{y}_0)$  in a closed form at an arbitrary timestep  $t$ :  $\mathbf{y}_t = \sqrt{\bar{\alpha}_t} \mathbf{y}_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon$ , where  $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ . The true posterior  $q(\mathbf{y}_{t-1} | \mathbf{y}_t)$  can be approximated by a deep neural network to predict the mean and variance of  $\mathbf{y}_{t-1}$  with the following parameterization,

$$\begin{aligned} p_\theta(\mathbf{y}_{t-1} | \mathbf{y}_t, \mathbf{x}_p, \mathbf{x}_s) &= \mathcal{N}(\mathbf{y}_{t-1}; \mu_\theta(\mathbf{y}_t, t, \mathbf{x}_p, \mathbf{x}_s), \\ &\quad \Sigma_\theta(\mathbf{y}_t, t, \mathbf{x}_p, \mathbf{x}_s)). \end{aligned} \quad (2)$$

**Noise prediction module  $\mathcal{H}_N$ :** Instead of directly deriving  $\mu_\theta$  following [6], we predict the noise  $\epsilon_\theta(\mathbf{y}_t, t, \mathbf{x}_p, \mathbf{x}_s)$  using our noise prediction module  $\mathcal{H}_N$ . The noisy image  $\mathbf{y}_t$

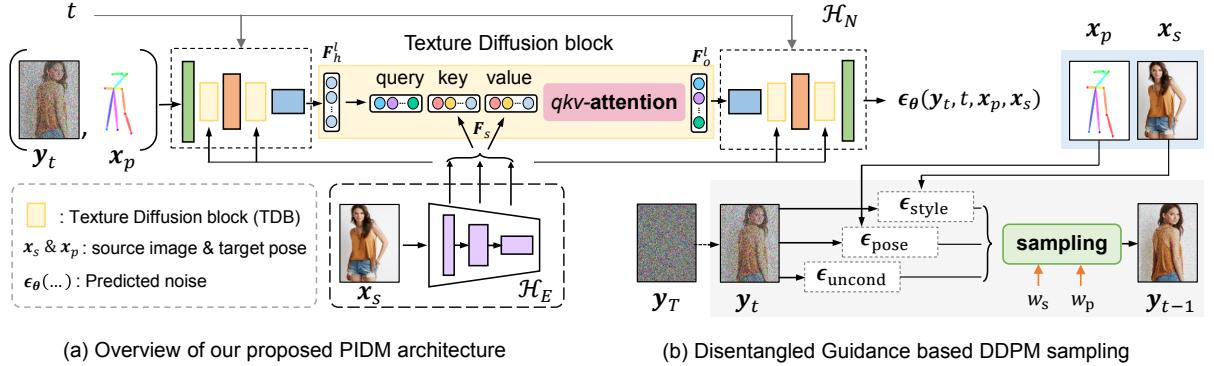


Figure 2. (a) The proposed PIDM framework is a UNet-based network composed of a noise prediction module  $\mathcal{H}_N$  and a texture encoder  $\mathcal{H}_E$ . The encoder  $\mathcal{H}_E$  encodes the texture patterns of the source image  $x_s$ . To obtain multi-scale features, we derive output from the different layers of  $\mathcal{H}_E$  resulting in a stacked feature representation  $F_s$ . To transfer rich multi-scale texture patterns from the source image distribution to the noise prediction module  $\mathcal{H}_N$ , we propose to use cross-attention based *Texture diffusion blocks* (TDB) that are embedded in different layers of  $\mathcal{H}_N$ . This allows the network to fully exploit the correspondences between source and target appearance, thus resulting in distortion-free images. (b) During inference, to amplify the conditional signal of  $x_s$  and  $x_p$  in the sampled images, we adapt the Classifier-free guidance [7] in our sampling technique to achieve *disentangled guidance*.

is concatenated with the target pose  $x_p$  and passed through  $\mathcal{H}_N$  to predict the noise.  $x_p$  will guide the denoising process and ensure that the intermediate noise representations and the final image follow the given skeleton structure. To inject the desired texture patterns into the noise predictor branch, we provide the multiscale features of the texture encoder  $\mathcal{H}_E$  through *Texture diffusion blocks* (TDB). To train the denoising process, we first generate a noisy sample  $y_t \sim q(y_t|y_0)$  by adding Gaussian noise  $\epsilon$  to  $y_0$ , then train a conditional denoising model  $\epsilon_\theta(y_t, t, x_p, x_s)$  to predict the added noise using a standard MSE loss:

$$L_{mse} = \mathbb{E}_{t \sim [1, T], y_0 \sim q(y_0), \epsilon} \|\epsilon - \epsilon_\theta(y_t, t, x_p, x_s)\|^2. \quad (3)$$

Nichol *et al.* [17] present an effective learning strategy as an improved version of DDPM with fewer steps needed and applies an additional loss term  $L_{vib}$  to learn the variance  $\Sigma_\theta$ . The overall hybrid objective that we adopt is as follows:

$$L_{hybrid} = L_{mse} + L_{vib}. \quad (4)$$

### 3.2. Texture Diffusion Blocks (TDB)

To mix the style of the source image within the noise prediction branch, we employ cross-attention based TDB units that are embedded in different layers of  $\mathcal{H}_N$ . Let  $F_h^l$  be the noise features in layer  $l$  of the noise prediction branch. Given the multiscale texture features  $F_s$  derived from the  $\mathcal{H}_E$  as input to TDB units, the attention module essentially computes the region of interest with respect to each query position, which is important to subsequently denoise the given noisy sample in the direction of the desired texture patterns. The keys  $K$  and values  $V$  are derived from  $\mathcal{H}_E$  while queries  $Q$  are obtained from noise features  $F_h^l$ . The

attention operation is formulated as follows:

$$\begin{aligned} Q &= \phi_q^l(F_h^l), \quad K = \phi_k^l(F_s), \quad V = \phi_v^l(F_s) \\ F_{att}^l &= \frac{QK^T}{\sqrt{C}}, \quad F_o^l = W^l \text{softmax}(F_{att}^l)V + F_h^l, \end{aligned} \quad (5)$$

where  $\phi_q^l$ ,  $\phi_k^l$ ,  $\phi_v^l$  are layer-specific  $1 \times 1$  convolution operators.  $W^l$  refers to learnable weights to generate final cross-attended features  $F_o^l$ . We adopt TDB for the feature at specific resolutions, *i.e.*,  $32 \times 32$ ,  $16 \times 16$ , and  $8 \times 8$ .

### 3.3. Disentangled Guidance based Sampling

Once the model learns the conditional distribution, inference is performed by first sampling a Gaussian noise  $y_T \sim \mathcal{N}(0, I)$  and then sampling from  $p_\theta(y_{t-1}|y_t, x_p, x_s)$ , from  $t = T$  to  $t = 1$  in an iterative manner. While the generated images using the vanilla sampling technique look photorealistic, they often do not strongly correlate with the conditional source image and target pose input. To amplify the effect of the conditioning signal  $x_s$  and  $x_p$  in the sampled images, we adapt Classifier-free guidance [7] in our multi-conditional sampling procedure. We observe that in order to sample images that not only fulfil the style requirement, but also ensure perfect alignment with the target pose input, it is important to employ *disentangled guidance* with respect to both style and pose. To enable disentangled guidance, we use the following equation:

$$\epsilon_{cond} = \epsilon_{uncond} + w_p \epsilon_{pose} + w_s \epsilon_{style}, \quad (6)$$

where  $\epsilon_{uncond} = \epsilon_\theta(y_t, t, \emptyset, \emptyset)$  is the unconditioned prediction of the model, where we replace both conditions with the all-zeros tensor  $\emptyset$ . The pose-guided prediction and the style-guided prediction are respectively represented by  $\epsilon_{pose} = \epsilon_\theta(y_t, t, x_p, \emptyset) - \epsilon_{uncond}$  and  $\epsilon_{style} =$



Figure 3. Qualitative comparisons with several state-of-the-art models on the DeepFashion dataset. The inputs to the model are the target pose  $\mathbf{x}_p$  and the source image  $\mathbf{x}_s$ . From left-to-right the results are of ADGAN [14], PISE [23], GFLA [19], DPTN [24], CASD [28], NTED [18] and ours respectively. (figure best viewed in zoom)

$\epsilon_\theta(\mathbf{y}_t, t, \emptyset, \mathbf{x}_s) - \epsilon_{\text{uncond.}}$   $w_p$  and  $w_s$  are guidance scale corresponding to pose and style. In practice, the diffusion model learns both conditioned and unconditioned distributions during training by randomly setting conditional variables  $\mathbf{x}_p$  and  $\mathbf{x}_s = \emptyset$  for  $\eta\%$  of the samples, so that  $\epsilon_\theta(\mathbf{y}_t, t, \emptyset, \emptyset)$  approximates  $p(\mathbf{y}_0)$  more faithfully.

## 4. Experiments

**Datasets:** We carry out experiments on DeepFashion In-shop Clothes Retrieval Benchmark [11] and Market-1501 [27] dataset. DeepFashion contains 52,712 high-resolution images of fashion models. Following the same data configuration in [30], we split this dataset into training and testing subsets with 101,966 and 8,570 pairs, respectively. Skeletons are extracted by OpenPose [1]. Market-1501 contains 32,668 low-resolution images. The images

vary in terms of the viewpoints, background, illumination, etc. For both datasets, personal identities of the training and testing sets do not overlap.

**Evaluation Metrics:** We evaluate the model using three different metrics. *Structure Similarity Index Measure* (SSIM) [22] and *Learned Perceptual Image Patch Similarity* (LPIPS) [26] are used to quantify the reconstruction accuracy. SSIM calculates the pixel-level image similarity, while LPIPS computes the distance between the generated images and reference images at the perceptual domain. *Frèchet Inception Distance* (FID) [5] is used to measure the realism of the generated images. It calculates the Wasserstein-2 distance between distributions of the generated images and the ground-truth images.

**Implementation Details:** Our PIDM model has been trained with  $T = 1000$  noising steps and a linear noise schedule. During training, we adopt an exponential mov-

Table 1. Quantitative Comparison of the proposed PIDM with several state-of-the-art models in terms of *Frèchet Inception Distance* (FID), *Structure Similarity Index Measure* (SSIM) and *Learned Perceptual Image Patch Similarity* (LPIPS). The results are shown on both  $256 \times 176$  and  $512 \times 352$  resolution for DeepFashion and  $128 \times 64$  resolution for Market-1501 dataset.

Dataset	Methods	FID( $\downarrow$ )	SSIM( $\uparrow$ )	LPIPS( $\downarrow$ )
DeepFashion [11] ( $256 \times 176$ )	Def-GAN [20]	18.457	0.6786	0.2330
	PATN [30]	20.751	0.6709	0.2562
	ADGAN [14]	14.458	0.6721	0.2283
	PISE [23]	13.610	0.6629	0.2059
	GFLA [19]	10.573	0.7074	0.2341
	DPTN [24]	11.387	0.7112	0.1931
	CASD [28]	11.373	0.7248	0.1936
	NTED [18]	8.6838	0.7182	0.1752
	<b>PIDM (Ours)</b>	<b>6.3671</b>	<b>0.7312</b>	<b>0.1678</b>
DeepFashion [11] ( $512 \times 352$ )	CocosNet2 [29]	13.325	0.7236	0.2265
	NTED [18]	7.7821	0.7376	0.1980
	<b>PIDM (Ours)</b>	<b>5.8365</b>	<b>0.7419</b>	<b>0.1768</b>
Market-1501 [27] ( $128 \times 64$ )	Def-GAN [20]	25.364	0.2683	0.2994
	PTN [30]	22.657	0.2821	0.3196
	GFLA [19]	19.751	0.2883	0.2817
	DPTN [24]	18.995	0.2854	0.2711
	<b>PIDM (Ours)</b>	<b>14.451</b>	<b>0.3054</b>	<b>0.2415</b>

ing average (EMA) of the denoising network weights with 0.9999 decay. In all experiments, we use a batch size of 8. Adam optimizer is used with learning rate set to  $2e^{-5}$ . For disentangled guidance, we use  $\eta = 10$ . For sampling, the values of  $w_p$  and  $w_s$  are set to 2.0. For the DeepFashion dataset, we train our model using  $256 \times 176$  and  $512 \times 352$  images. For Market-1501, we use  $128 \times 64$  images.

#### 4.1. Quantitative and Qualitative Comparisons

We quantitatively compare (Tab. 1) our proposed PIDM with several state-of-the-art methods, including Def-GAN [20], PATN [30], ADGAN [14], PISE [23], GFLA [19], DPTN [24], CASD [28], CocosNet2 [29] and NTED [18]. The experiments are done on both  $256 \times 176$  and  $512 \times 352$  resolution for DeepFashion and  $128 \times 64$  resolution for Market-1501 dataset. Tab. 1 shows that our model achieves the best FID score indicating that our model can generate higher-quality images compared to the previous approaches. Furthermore, PIDM performs favorably against other methods in terms of the reconstruction metrics SSIM and LPIPS. This means that our model can generate images with not only accurate structures, but also can correctly transfer the texture of the source image to the target pose.

In Fig. 3, we present a comprehensive visual comparison of our method with other state-of-the-art frameworks on DeepFashion dataset<sup>1</sup>. The results of the baselines are obtained using pre-trained models provided by the corre-

<sup>1</sup>Additional results are provided in supplementary material.

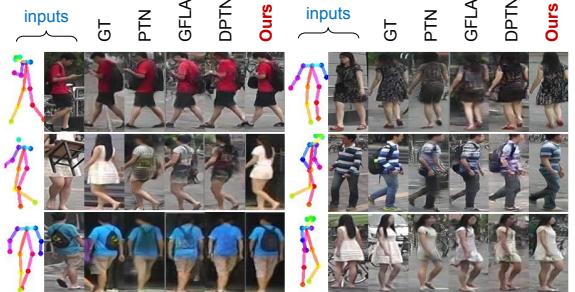


Figure 4. Qualitative comparisons with several state-of-the-art models on the Market-1501 dataset. From left-to-right the results are of PTN [30], GFLA [19], DPTN [24] and ours respectively.

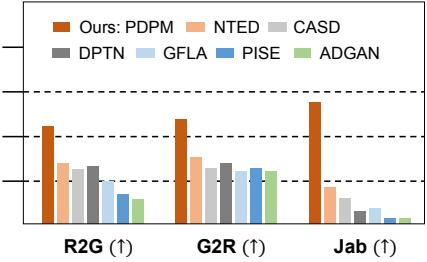


Figure 5. User study results on DeepFashion dataset in terms of R2G, G2R and Jab metric. Higher values indicate PIDM is preferred more often over the compared approaches.

sponding authors. It can be observed that PISE [23] and ADGAN [14] fail to generate sharp images and cannot keep the consistency of shape and texture. While GFLA [19] somewhat preserves the texture in the source image, it struggles to obtain reasonable results for the invisible regions of the source image (e.g., 4~5-th rows in the left column). NTED [18] and CASD [28] improve the results slightly but they are still not able to adequately preserve the source appearance in complex scenarios (e.g., 1~3-th rows in the left and 7~10-th rows in the right column). In comparison, our proposed PIDM accurately retains the appearance of the source while also producing images that are more natural and sharper. Moreover, even if the target pose is complex (e.g., 10-th row in the left column), our method can still generate it precisely. We also visually compare our method with other baselines on the Market-1501 dataset in Fig. 4. Even with the complex background, PIDM still performs favorably in generating photo-realistic results as well as a consistent appearance with the source image.

#### 4.2. User Study

To demonstrate the effectiveness of our model in terms of human perception, we present our user study on 100 human participants. The user study is conducted to measure different aspects of our model with respect to ground truth images and generated images using other methods. (i) To

Table 2. Impact of our texture diffusion blocks (TDB) and disentangled classifier-free (DCF) guidance in our proposed diffusion model on DeepFashion dataset.

Methods	FID( $\downarrow$ )	SSIM( $\uparrow$ )	LPIPS( $\downarrow$ )
B1: Baseline $^\dagger$ (concat)	10.813	0.6911	0.2112
B2: Baseline	9.8510	0.7005	0.1983
B3: Baseline + TDB	7.5133	0.7178	0.1870
B4: Baseline + TDB + CF-guidance	6.8176	0.7195	0.1769
<b>Ours: Baseline + TDB + DCF-guidance</b>	<b>6.3671</b>	<b>0.7312</b>	<b>0.1676</b>

compare with ground-truth images, we randomly select 30 generated images and 30 real images from the test set. Participants are required to determine whether a given image is real or fake. Following [28, 30], we adopt two metrics: *R2G* and *G2R*. *R2G* is the percentage of the real images classified as generated images and *G2R* is the percentage of the generated images classified as real images. **(ii)** To compare with other methods, we randomly select 30 sets of images where each set includes source image, target pose, ground-truth and images generated by our method and the baseline. The participants are required to select the best image with respect to the provided source image and ground truth. The participants are advised to provide their response based on the ability of each competing approach towards producing accurate texture patterns and pose structure. We quantify this using another metric called *Jab* which is defined as the percentage of images considered the best among all models. Higher values of these three metrics mean better performance. The result of the study is shown in Fig. 5. PIDM performs favorably against all baselines for all three metrics on DeepFashion dataset. For instance, PIDM images were interpreted as real images 48% (*G2R*) out of total cases, which is nearly 18% higher than the second best model. Our *Jab* score is 56% showing that the participants favor our approach more frequently than other methods.

### 4.3. Ablation Study

We perform multiple ablation studies to validate the merits of the proposed contributions. Tab. 2 shows the impact of texture diffusion blocks (TDB) and distangled classifier-free (DCF) guidance on the DeepFashion dataset. Baseline $^\dagger$  neither employs a separate texture encoder nor uses TDB units. It only comprises of a UNet-based noise prediction module where the target pose and source image are concatenated with the noisy image and passed through the module to output a denoised version of the image. We extend the Baseline $^\dagger$  with additional texture encoder to extract meaningful texture representations from the source image which are concatenated with respective layers of the noise prediction module. This is denoted as Baseline in Tab. 2. While the Baseline is able to generate realistic person images, it has a limited ability to retain the appearance of the source image as shown in Fig. 6. To effectively model the complex interplay between appearance

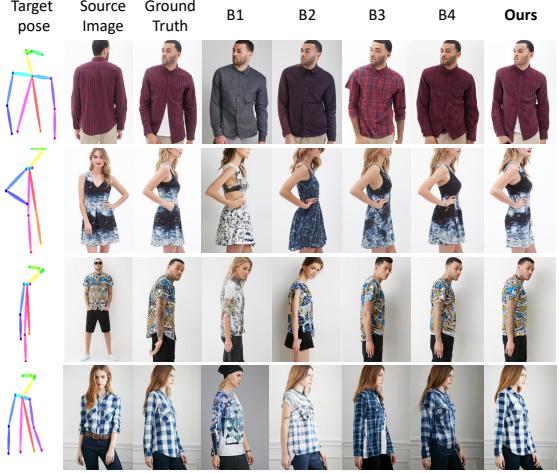


Figure 6. Qualitative ablation results on the DeepFashion dataset. The images in this figure correspond to the ablation studies in Tab. 2. Compared to the B1 (Baseline $^\dagger$ ), the B2 (Baseline) benefits from an additional texture encoder improving the correlation with the source appearance. The B3 (Baseline+TDB) further improves the results using texture diffusion module (TDB) that effectively model the complex interplay between appearance and pose information. Our final proposed PIDM referred as Baseline+TDB+DCF-guidance employs the disentangled classifier-free (DCF) guidance. Contrary to the B4 (Baseline+TDB+CF-guidance), our proposed PIDM employs *disentangled guidance* with respect to both style and pose in order to tightly align the output image style and pose with the source image appearance and target pose.

and pose information, we integrate texture diffusion module (TDB) into the noise prediction module. We refer to this as Baseline+TDB. On the DeepFashion dataset, in comparison to the Baseline, the Baseline+TDB improves the FID by a margin of 2.3377. While the generated images using the vanilla sampling technique look photo-realistic, they often do not strongly correlate with the conditional source image and target pose input. To improve the correlation, we first use vanilla classifier-free (CF) guidance during the sampling. While the vanilla CF guidance slightly improves the FID score, we observe that in order to sample high-quality images that not only fulfil the style requirement, but also ensure perfect alignment with the target pose input, it is important to employ disentangled guidance with respect to both style and pose. We here refer to our final proposed approach Baseline+TDB+DCF-guidance where we employ disentangled classifier-free (DCF) guidance to tightly align the output image style and pose with the source image appearance and target pose. In comparison to the Baseline+TDB, our proposed DCF guidance improves the FID, SSIM and LPIPS by a margin of 1.1462, 0.0134 and 0.0194 respectively.



Figure 7. (a) Qualitative evaluation of our proposed PIDM for appearance control. The results demonstrate the seamless editing capabilities of our model. Images are generated by controlling the appearance of the reference image while maintaining the person’s pose and identity. For each example, the first row contains the style images and the second row contains the generated images. (b) The interpolation results show that the clothes’ texture gradually changes from the style of the left image to that of the right image. (*figure best viewed in zoom*)

#### 4.4. Appearance Control and Editing

Our proposed PIDM inherits the flexibility and controllability of diffusion models that enable appearance control by combining cloth textures extracted from style images into the reference image. Given a source (style) image  $\mathbf{x}_s$ , a reference image  $\mathbf{y}^{ref}$ , and a binary mask  $\mathbf{m}$  that marks the region of interest in the reference image, the problem is to generate an image  $\bar{\mathbf{y}}$ , s.t. the appearance of  $\bar{\mathbf{y}} \odot \mathbf{m}$  is consistent with the source image  $\mathbf{x}_s$ , while the complementary area remains same. To achieve this, we first calculate  $\mathbf{y}_t^{ref}$  in a time step  $t$ :  $\mathbf{y}_t^{ref} = \sqrt{\alpha_t} \mathbf{y}^{ref} + \sqrt{1 - \alpha_t} \epsilon$ , where  $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ . During inference, starting from a Gaussian noise  $\mathbf{y}_T \sim \mathcal{N}(0, \mathbf{I})$ , we predict  $\mathbf{y}_t$  iteratively using the trained diffusion model from  $t = T$  to  $t = 1$ . In each step  $t$ , we use the binary mask  $\mathbf{m}$  to retain the unmasked regions of  $\mathbf{y}^{ref}$  by using the relation:  $\mathbf{y}_t = \mathbf{m} \odot \mathbf{y}_t + (1 - \mathbf{m}) \odot \mathbf{y}_t^{ref}$ . The results of the appearance control task are shown in Fig. 7(a). We observe that our model can seamlessly combine the areas of interest and generate coherent images with realistic textures.

**Style Interpolation:** In Fig. 7(b), we show our interpolation results between two style images. We use DDIM [21] sampling to enable smooth interpolation using our trained diffusion model. Specifically, we use spherical linear interpolation between noises  $\mathbf{y}_T^1$  and  $\mathbf{y}_T^2$ , and linear interpolation between style features  $\mathbf{F}_s^1$  and  $\mathbf{F}_s^2$ . As shown in Fig. 7(b), the texture of the clothes gradually changes from the style of the left source image to that of the right source image.

#### 4.5. Application to Person Re-identification

Here, we evaluate the applicability of the images generated by our PIDM as a source of data augmentation for the downstream task of person re-identification (re-ID). We perform the re-ID experiment on Market-1501 dataset. We randomly select 20%, 40%, 60%, and 80% of total training set of real Market-1501 dataset such that at least one image per identity is selected. We denote the obtained set as  $\mathbb{D}_{tr}$ . We

Table 3. The person re-ID results on ResNet50 backbone in terms of mAP scores using images generated by different methods. The right-most column shows results when 30K generated images are added to the real set. *Standard* means no augmentation is used.

Methods	Percentage of real images				100%(+30K)
	20%	40%	60%	80%	
Standard	33.4	56.6	64.9	69.2	76.7
PTN [30]	55.6	57.3	67.1	72.5	76.8
GFLA [19]	57.3	59.7	67.6	73.2	76.8
DPTN [24]	58.1	62.6	69.0	74.2	77.1
PIDM (Ours)	<b>61.3</b>	<b>64.8</b>	<b>71.6</b>	<b>75.3</b>	<b>78.4</b>

first initialize a ResNet50 backbone network using  $\mathbb{D}_{tr}$  for the re-ID task. We refer to this as *Standard* in Tab. 3. Then, we augment  $\mathbb{D}_{tr}$  with images generated by our PIDM. The images are generated using randomly chosen image of the same identity in  $\mathbb{D}_{tr}$  as the source image. The target pose is randomly selected from  $\mathbb{D}_{tr}$ . Consequently, an augmented training set  $\mathbb{D}_{aug}$  is created from all these generated images and the real set  $\mathbb{D}_{tr}$ . We finetune the ResNet50 backbone using the augmentation set  $\mathbb{D}_{aug}$ . As shown in Tab. 3, PIDM achieves consistent improvements over previous works.

## 5. Conclusion

We proposed a diffusion-based approach, PIDM, for pose-guided person image generation. PIDM disintegrates the complex transfer problem into a series of simpler forward-backward denoising steps. This helps in learning plausible source-to-target transformation trajectories that result in faithful textures and undistorted appearance details. We introduce *texture diffusion module* and *disentangled classifier-free guidance* to accurately model the correspondences between appearance and pose information available in source and target images. We show the effectiveness of PIDM on two datasets by performing extensive qualitative, quantitative and human-based evaluations. In addition, we show how our generated images can help in downstream tasks such as person re-identification.

## References

- [1] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017. 5
- [2] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *NeurIPS*, 2021. 2
- [3] Patrick Esser, Ekaterina Sutter, and Björn Ommer. A variational u-net for conditional appearance and shape generation. In *CVPR*, 2018. 2
- [4] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. In *Communications of the ACM*, 2020. 1
- [5] 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. 5
- [6] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 2, 3
- [7] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 2, 3, 4
- [8] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 1
- [9] Yining Li, Chen Huang, and Chen Change Loy. Dense intrinsic appearance flow for human pose transfer. In *CVPR*, 2019. 2
- [10] Wen Liu, Zhixin Piao, Jie Min, Wenhan Luo, Lin Ma, and Shenghua Gao. Liquid warping gan: A unified framework for human motion imitation, appearance transfer and novel view synthesis. In *ICCV*, 2019. 2
- [11] 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, 5, 6
- [12] 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, 3
- [13] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. Pose guided person image generation. In *NeurIPS*, 2017. 2
- [14] Yifang Men, Yiming Mao, Yuning Jiang, Wei-Ying Ma, and Zhouhui Lian. Controllable person image synthesis with attribute-decomposed gan. In *CVPR*, 2020. 2, 3, 5, 6
- [15] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. 2
- [16] Alex 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. *arXiv preprint arXiv:2112.10741*, 2021. 2
- [17] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *ICML*, 2021. 4
- [18] 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, 3, 5, 6
- [19] 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, 3, 5, 6, 8
- [20] Aliaksandr Siarohin, Enver Sangineto, Stéphane Lathuilière, and Nicu Sebe. Deformable gans for pose-based human image generation. In *CVPR*, 2018. 2, 6
- [21] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021. 8
- [22] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. In *TIP*, 2004. 5
- [23] JinSong Zhang, Kun Li, Yu-Kun Lai, and Jingyu Yang. Pise: Person image synthesis and editing with decoupled gan. In *CVPR*, 2021. 1, 2, 3, 5, 6
- [24] Pengze Zhang, Lingxiao Yang, Jian-Huang Lai, and Xiaohua Xie. Exploring dual-task correlation for pose guided person image generation. In *CVPR*, 2022. 5, 6, 8
- [25] Pan Zhang, Bo Zhang, Dong Chen, Lu Yuan, and Fang Wen. Cross-domain correspondence learning for exemplar-based image translation. In *CVPR*, 2020. 2, 3
- [26] 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. 5
- [27] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *ICCV*, 2015. 2, 5, 6
- [28] 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, 3, 5, 6, 7
- [29] 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, 6
- [30] Zhen Zhu, Tengteng Huang, Baoguang Shi, Miao Yu, Bofei Wang, and Xiang Bai. Progressive pose attention transfer for person image generation. In *CVPR*, 2019. 1, 2, 5, 6, 7, 8