Person Image Synthesis via Denoising Diffusion Model

Manu S Pillai, Prudvi Kamtam, Mukund Dhar and Adeel Yousaf
Contents

• Problem Formulation
• Existing Points
• Diffusion based Pose Synthesis
• Contributions
• Proposed Model
• Experimentation Details
• Ablation Study
• Appearance Control
• Person Re-Identification
• Demo
Problem Formulation
Problem Formulation
Problem Formulation

Pose  Style
Problem Formulation

Pose

Style

Picture of a floral patterned coat
Existing Methods

GANs

Discriminator D

Generator $G$

single forward pass
Diffusion based Pose Synthesis

$x_s$
Diffusion based Pose Synthesis

$x_s$

$x_p$
Diffusion based Pose Synthesis

\[ x_s \]

\[ x_p \]

\[ y_T \sim \mathcal{N}(0, I) \]

\[ p_\theta(y | x_s, x_p) \]

\[ T \text{ Diffusion steps} \]
Diffusion based Pose Synthesis

\[ y_T \sim \mathcal{N}(0, I) \]

T Diffusion steps

\[ p_\theta(y | x_s, x_p) \]
Contributions

- First diffusion-based approach for pose guided person synthesis task
- New state of the art on DeepFashion and Market-1501.
- Synthesized images can be used to improve performance in downstream tasks
Overall framework
Overall framework

Noise prediction module ($H_N$)
Overall framework

\[ y_t \quad x_p \quad \xrightarrow{\text{Noise prediction module (} H_N \text{)}} \]

\[ x_s \]
Overall framework

Noise prediction module ($H_N$)

texture encoder ($H_E$)

$y_t$ $x_p$ $x_S$
Overall framework

Noise prediction module ($H_N$)

$x_p$ and $x_s$ will guide the denoising process
Proposed Method

$y_t$

Forward Diffusion Process

$y_{t-1}$

Backward Diffusion Process

$y_t$
Proposed Method

Backward Diffusion Process

\[ y_t \]

\[ x_p \]

\[ x_S \]

\[ y_{t-1} \]
Backward Diffusion Process
Person Image Synthesis via DDM
Person Image Synthesis via DDM

\[
x_p, y_t \\
\xrightarrow{U_{net}^{enc}} U_{net}^{dec} \xrightarrow{} y_{t-1}
\]
Person Image Synthesis via DDM

$x_p$ \rightarrow \overrightarrow{U_{\text{enc}}_{\text{net}}} \rightarrow \text{Texture Diffusion block} \rightarrow \overleftarrow{U_{\text{dec}}_{\text{net}}} \rightarrow y_{t-1}$
Texture Diffusion block
Texture Diffusion block

$x_p$

$y_t$

$U_{enc}^{net}$

Texture Diffusion block

$U_{dec}^{net}$

$y_{t-1}$

$x_s$
Texture Diffusion block
Texture Diffusion block

\[ x_p \rightarrow U_{\text{enc}}^\text{net} \rightarrow F_h^l \rightarrow U_{\text{net}} \rightarrow y_{t-1} \]

\[ x_s \rightarrow F_S \rightarrow \mathcal{H}_E \]
Texture Diffusion block

$x_p$ → $U_{enc}^{\text{net}}$ → $F_h^l$ → $U_{dec}^{\text{net}}$ → $y_{t-1}$
Texture Diffusion block

\[ x_p \rightarrow U_{\text{enc}} \rightarrow F_h^l \rightarrow \text{Texture Diffusion block} \rightarrow \rightarrow U_{\text{dec}} \rightarrow y_{t-1} \]
Texture Diffusion block

$x_p \rightarrow U_{\text{net}}^{\text{enc}} \rightarrow F_h^l \rightarrow \text{Texture Diffusion block} \rightarrow F_o^l \rightarrow U_{\text{net}}^{\text{dec}} \rightarrow y_{t-1}$
Texture Diffusion block

\[ y_{t-1} = \epsilon_{\theta}(y_t, t, x_p, x_s) \]
Sampling
Vanilla Sampling

\[ y_T \sim \mathcal{N}(0, I) \]
Vanilla Sampling

\[ y_T \sim N(0, I) \]

\[ \mathcal{H}_N \]

\[ \varepsilon_{\theta} (y_{t-1} | y_t, x_p, x_s) \]
Vanilla Sampling

\[ x_p \]

\[ y_T \sim \mathcal{N}(0, I) \]

\[ \mathcal{H}_N \]

\[ \mathcal{H}_E \]

\[ x_s \]

\[ \varepsilon_\theta (y_{t-1} | y_t, x_p, x_s) \]

sampling

\[ y_{t-1} \]
Vanilla Sampling

\[ x_p \quad x_s \]

GT

Output
Vanilla Sampling

\[ x_p \quad x_s \quad GT \quad Output \]

Not completely aligned with style or pose

Solution? Classifier free Guidance
Classifier free Guidance

\[ y_T \sim \mathcal{N}(0, I) \]

\[ x_p, y_T \]

\[ \mathcal{H}_N \]

\[ \varepsilon_\theta(y_t, t, x_p, x_s) \]

\[ \mathcal{H}_E \]

\[ x_s \]

\[ y_t \]
Classifier free Guidance

\[ y_T \sim \mathcal{N}(0, I) \]

\[ \epsilon_\theta(y_t, t, \emptyset, \emptyset) \]
Classifier Free Guidance

\[ y_T \sim \mathcal{N}(0, I) \]

\[ \mathcal{H}_N \]

\[ \hat{\epsilon}_{\text{cond}} = \epsilon_{\text{uncond}} + S \cdot (\epsilon_{\text{cond}} - \epsilon_{\text{uncond}}) \]

\[ \epsilon_{\text{uncond}} = \epsilon_{\theta} (y_{t-1} \mid y_t, \emptyset, \emptyset) \]

\[ \epsilon_{\text{cond}} = \epsilon_{\theta} (y_{t-1} \mid y_t, x_p, x_s) \]
Classifier Free Guidance

\[ y_T \sim \mathcal{N}(0, I) \]

\[ y_T \to \mathcal{H}_N \to \mathcal{H}_E \to x_s \]

\[ \mathcal{H}_N \to \mathcal{E}_{\text{cond}} \to \text{sampling} \to y_{t-1} \]
Classifier free Guidance

$x_p$  $x_S$  GT
Classifier free Guidance

Still Not completely aligned with BOTH style and pose

Solution? Disentangled Guidance based Sampling
Disentangled Guidance based Sampling

\[ \epsilon_{\text{cond}} = \epsilon_{\text{uncond}} + w_p \epsilon_{\text{pose}} + w_s \epsilon_{\text{style}} \]

\[ = \epsilon_\theta(y_t, t, \emptyset, \emptyset) \]

\[ = \epsilon_\theta(y_t, t, x_p, \emptyset) - \epsilon_{\text{uncond}} \]

\[ = \epsilon_\theta(y_t, t, \emptyset, x_s) - \epsilon_{\text{uncond}}. \]
Disentangled Guidance based Sampling

\[ \epsilon_{\text{cond}} = \epsilon_{\text{uncond}} + w_p \epsilon_{\text{pose}} + w_s \epsilon_{\text{style}} \]
Experimentation Details: Datasets

DeepFashion In-shop Clothes Retrieval
- 52,712 high-resolution images of fashion models

Market- 1501
- 32,668 low-resolution images
Experimentation Details: Datasets

DeepFashion In-shop Clothes Retrieval

Market- 1501
1. **Structure Similarity Index Measure (SSIM)**
SSIM calculates the pixel-level image similarity

2. **Learned Perceptual Image Patch Similarity (LPIPS)**
LPIPS computes the distance between the generated images and reference images at the perceptual domain.

3. **Fréchet Inception Distance (FID)**
FID is used to measure the photo-realism of the generated images
## Quantitative Comparisons

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Methods</th>
<th>FID(↓)</th>
<th>SSIM(↑)</th>
<th>LPIPS(↓)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PATN [30]</td>
<td>20.751</td>
<td>0.6709</td>
<td>0.2562</td>
</tr>
<tr>
<td></td>
<td>ADGAN [14]</td>
<td>14.458</td>
<td>0.6721</td>
<td>0.2283</td>
</tr>
<tr>
<td></td>
<td>PISE [23]</td>
<td>13.610</td>
<td>0.6629</td>
<td>0.2059</td>
</tr>
<tr>
<td></td>
<td>GFLA [19]</td>
<td>10.573</td>
<td>0.7074</td>
<td>0.2341</td>
</tr>
<tr>
<td></td>
<td>DPTN [24]</td>
<td>11.387</td>
<td>0.7112</td>
<td>0.1931</td>
</tr>
<tr>
<td></td>
<td>CASD [28]</td>
<td>11.373</td>
<td>0.7248</td>
<td>0.1936</td>
</tr>
<tr>
<td></td>
<td>NTED [18]</td>
<td>8.6838</td>
<td>0.7182</td>
<td>0.1752</td>
</tr>
<tr>
<td></td>
<td>PIDM (Ours)</td>
<td><strong>6.3671</strong></td>
<td><strong>0.7312</strong></td>
<td><strong>0.1678</strong></td>
</tr>
<tr>
<td>(256 × 176)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>NTED [18]</td>
<td>7.7821</td>
<td>0.7376</td>
<td>0.1980</td>
</tr>
<tr>
<td></td>
<td>PIDM (Ours)</td>
<td><strong>5.8365</strong></td>
<td><strong>0.7419</strong></td>
<td><strong>0.1768</strong></td>
</tr>
<tr>
<td>(512 × 352)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Market-1501 [27]</td>
<td>Def-GAN [20]</td>
<td>25.364</td>
<td>0.2683</td>
<td>0.2994</td>
</tr>
<tr>
<td></td>
<td>PTN [30]</td>
<td>22.657</td>
<td>0.2821</td>
<td>0.3196</td>
</tr>
<tr>
<td></td>
<td>GFLA [19]</td>
<td>19.751</td>
<td>0.2883</td>
<td>0.2817</td>
</tr>
<tr>
<td></td>
<td>DPTN [24]</td>
<td>18.995</td>
<td>0.2854</td>
<td>0.2711</td>
</tr>
<tr>
<td></td>
<td>PIDM (Ours)</td>
<td><strong>14.451</strong></td>
<td><strong>0.3054</strong></td>
<td><strong>0.2415</strong></td>
</tr>
</tbody>
</table>
Qualitative Comparisons
Qualitative Comparisons
Qualitative Comparisons
Qualitative Comparisons
Qualitative Comparisons

NTED (CVPR’22)  PIDM (Ours)  NTED (CVPR’22)  PIDM (Ours)
Results

Results using images from fashion e-commerce site as source
## Results

Results comparing with several state-of-the-art on the Market-1501 dataset.
Ablation Study [B1: Baseline (concat)]

\[ x_p \]  \rightarrow  \[ U_{net}^{enc} \]  \rightarrow  \[ U_{net}^{dec} \]  \rightarrow  \[ y_{t-1} \]

\( x_s \)
Ablation Study
Ablation Study [B2: Baseline]
Ablation Study
Ablation Study [B3: Baseline + TDB]
Ablation Study

B4: Baseline + TDB + CF-guidance

Ours: Baseline + TDB + DCF-guidance
## Ablation Study

<table>
<thead>
<tr>
<th>Methods</th>
<th>FID (↓)</th>
<th>SSIM (↑)</th>
<th>LPIPS (↓)</th>
</tr>
</thead>
<tbody>
<tr>
<td>B1: Baseline † (concat)</td>
<td>10.813</td>
<td>0.6911</td>
<td>0.2112</td>
</tr>
<tr>
<td>B2: Baseline</td>
<td>9.8510</td>
<td>0.7005</td>
<td>0.1983</td>
</tr>
<tr>
<td>B3: Baseline + TDB</td>
<td>7.5133</td>
<td>0.7178</td>
<td>0.1870</td>
</tr>
<tr>
<td>B4: Baseline + TDB + CF-guidance</td>
<td>6.8176</td>
<td>0.7195</td>
<td>0.1769</td>
</tr>
<tr>
<td><strong>Ours: Baseline + TDB + DCF-guidance</strong></td>
<td><strong>6.3671</strong></td>
<td><strong>0.7312</strong></td>
<td><strong>0.1676</strong></td>
</tr>
</tbody>
</table>
Appearance Control

• The PIDM model inherits the flexibility and controllability of diffusion models that enable appearance control.
  • It is done by combining cloth textures extracted from style images into the reference image.

• The model can seamlessly combine the areas of interest and generate coherent output images with realistic textures.

![Qualitative evaluation of the proposed PIDM for appearance control.](image)
Appearance Control

1. Calculate \( y_t^{\text{ref}} = \sqrt{\alpha_t} y_t^{\text{ref}} + \sqrt{1 - \alpha_t} \varepsilon \) at a given time \( t \).
2. Predict \( y_t \) iteratively from \( t = T \) to \( t = 1 \) during inference.
3. In each step \( t \), use the binary mask \( m \) to retain \( y_t^{\text{ref}} \).

\[
\begin{align*}
\varepsilon & \sim \mathcal{N}(0, I) \\
y_T & \sim \mathcal{N}(0, I) \\
y_t &= m \odot y_t + (1 - m) \odot y_t^{\text{ref}}
\end{align*}
\]
Appearance Control

Style

Ref.
Person Re-Identification

- It aims at retrieving a person of interest across multiple non-overlapping cameras.
  - To determine whether a person-of-interest has appeared in another place at a distinct time captured by the same or different camera.

- The images generated by the PIDM as a source of data augmentation can be utilized for improved results.

- The images are augmented with the *Standard* images to finetune the ResNet50 backbone to perform Re-ID.
  - Randomly selected images from total training set of real Market-1501 dataset initialize the ResNet50 network - > *Standard*
Person Re-Identification

<table>
<thead>
<tr>
<th>Methods</th>
<th>Percentage of real images</th>
<th>100% (+30K)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>20%</td>
<td>40%</td>
</tr>
<tr>
<td>Standard</td>
<td>33.4</td>
<td>56.6</td>
</tr>
<tr>
<td>PTN [30]</td>
<td>55.6</td>
<td>57.3</td>
</tr>
<tr>
<td>GFLA [19]</td>
<td>57.3</td>
<td>59.7</td>
</tr>
<tr>
<td>DPTN [24]</td>
<td>58.1</td>
<td>62.6</td>
</tr>
<tr>
<td>PIDM (Ours)</td>
<td>61.3</td>
<td>64.8</td>
</tr>
</tbody>
</table>

The person re-ID results in terms of mAP scores.
Conclusion

• A diffusion-based approach for pose-guided person image generation.

• A texture diffusion module and a disentangled classifier-free guidance are introduced.
  • Helps in modeling the correspondences between appearance and pose information available in source and target images.

• The PIDM model is evaluated performing extensive qualitative and quantitative comparisons.
  • It performs well on metrics such as FID, SSIM and LPIPS.

• The model, in addition, can effectively help in downstream tasks such as person re-identification.
Demo
Demo

timestep = 999, T=1000
Demo
Demo
Thank you.