上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

SJTU

PhysDiff-VTON: Cross-Domain Physics Modeling and Trajectory Optimization for Virtual Try-On

Shibin Mei, Bingbing Ni,

2025/10/19

Paper ID: 12374

饮 水 思 源 • 爱 国 荣 校

01 Motivation

02 Method

03 Experiments

04 Conclusion

Motivation

The Dual Challenges of Virtual Try-On

Geometric Implausibility:

- Rigid warping (e.g., TPS) fails to model complex fabric dynamics.
- Results in unnatural wrinkles and stretching.

Texture Erosion:

- Iterative denoising in diffusion models smears high-frequency details
- Loss of patterns, text, and fine textures

badly warped garment

text becomes blurry

Motivation

Our Contributions: A Physics-Integrated Framework

- **Physics-Aware Deformation** : A pose-guided deformable warping module that simulates strain-dependent fabric dynamics.
- **Spectral Fidelity Preservation** : A wavelet-enhanced feature decomposition to explicitly preserve high-frequency texture details.
- **Trajectory Optimization (PRPO) :** A novel sampling strategy that optimizes the denoising path for spatiotemporal coherence.

The PhysDiff-VTON Framework

Inputs: Person Image (I^p), Garment Image (I^g), Pose Map (P)

Three Core Modules:

- Pose-Guided Deformable Warping
- Wavelet-Enhanced UNet
- PRPO Sampler

Pose-Guided Deformable Warping：Simulating Fabric Dynamics

- Predict spatial offsets conditioned on pose semantics.
- Offset prediction:

$$\Delta p = \mathcal{G}(\mathbf{F}_g \oplus (\mathbf{A} \odot \mathbf{E}_p))$$

- Deformable convolution:

$$\tilde{\mathbf{F}}_g(x) = \sum_{k=1}^{K} w_k \cdot \mathbf{F}_g(x + p_k + \Delta p_k)$$

Method

Wavelet-Enhanced Texture Preservation

- Explicitly decompose and reinforce frequency components.
- Wavelet transformation:

$$\{\mathbf{LL}, \mathbf{LH}, \mathbf{HL}, \mathbf{HH}\} = \mathrm{DWT}(\mathbf{X})$$

- Frequency-Gated Attention:

$$\mathbf{A} = \sigma(\mathbf{Conv}_1(\mathrm{GN}(\mathbf{Conv}_2(\mathrm{Concat}(\mathbf{LL}, \mathbf{LH}, \mathbf{HL}, \mathbf{HH})))))$$

- Final Feature:

$$\tilde{\mathbf{X}} = \mathbf{X} + \mathbf{A} \odot \mathbf{X}$$

Method

Trajectory Optimization (PRPO)

- Reformulate sampling as a variational problem minimizing an "Action Functional".

$$\mathcal{S}[x(t)] = \underbrace{\int_0^T \|s_\theta(x(t),t) - \nabla_x \log q_t(x(t))\|^2 dt}_{\text{Dynamic Matching}} + \lambda \underbrace{\int_0^T E(x(t))dt}_{\text{Potential Regularization}} + \sigma \underbrace{\int_0^T \|\xi(t)\|^2 dt}_{\text{Stochastic Control}},$$

- Three Potentials:

✓ Temporal Smoothness: For coherent wrinkle formation.

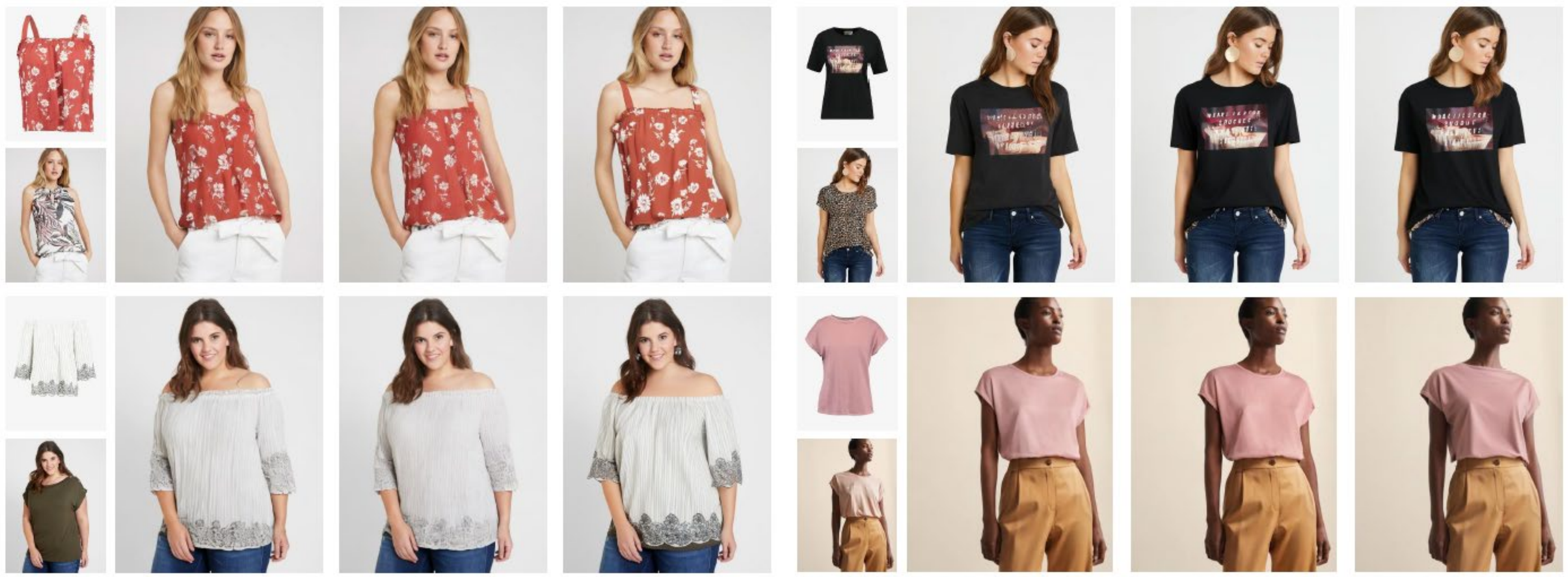
✓ Spatial TV: For piecewise smoothness and sharp edges.

✓ Multi-scale Similarity: For structural consistency.

Experiments

Dataset	VITON-HD				DressCode			
Method	LPIPS ↓	SSIM ↑	FID ↓	CLIP-I ↑	LPIPS ↓	SSIM ↑	FID ↓	CLIP-I ↑
GAN-based methods								
HR-VITON	0.115	0.883	9.70	0.832	0.112	0.910	21.42	0.771
GP-VTON	0.105	**0.898**	6.43	0.874	0.484	0.780	55.21	0.628
Diffusion-based methods								
LaDI-VTON	0.156	0.872	8.85	0.834	0.149	0.905	16.54	0.803
DCI-VTON	0.166	0.856	8.73	0.840	0.162	0.893	17.63	0.777
StableVITON	0.133	0.885	6.52	0.871	0.107	0.910	14.37	0.866
IDM-VITON	0.102	0.870	6.29	0.883	0.062	0.920	8.64	0.904
PhysDiff(Ours)	**0.093**	0.881	**6.21**	**0.894**	**0.055**	**0.932**	**8.27**	**0.918**

Experiments



Dataset	VITON-HD				DressCode			
Method	LPIPS ↓	SSIM ↑	FID ↓	CLIP-I ↑	LPIPS ↓	SSIM ↑	FID ↓	CLIP-I ↑
w/o Deform	0.102	0.873	6.33	0.882	0.069	0.922	8.51	0.898
w/o Wavelet	0.098	0.868	6.24	0.888	0.062	0.916	8.40	0.904
w/o PRPO	0.096	0.875	6.25	0.890	0.059	0.925	8.33	0.912
PhysDiff-VITON	**0.093**	**0.881**	**6.21**	**0.894**	**0.055**	**0.932**	**8.27**	**0.918**

Conclusion

Contribution

- Physics-aware deformation via pose-guided warping.
- Texture fidelity via wavelet-enhanced synthesis.
- Coherent generation via potential-regularized path optimization.

Future Work

- Extend to video try-on for temporal consistency.
- Model a wider range of fabric materials dynamically.

Thank you!

饮 水 思 源　爱 国 荣 校