南京航空航天大學
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

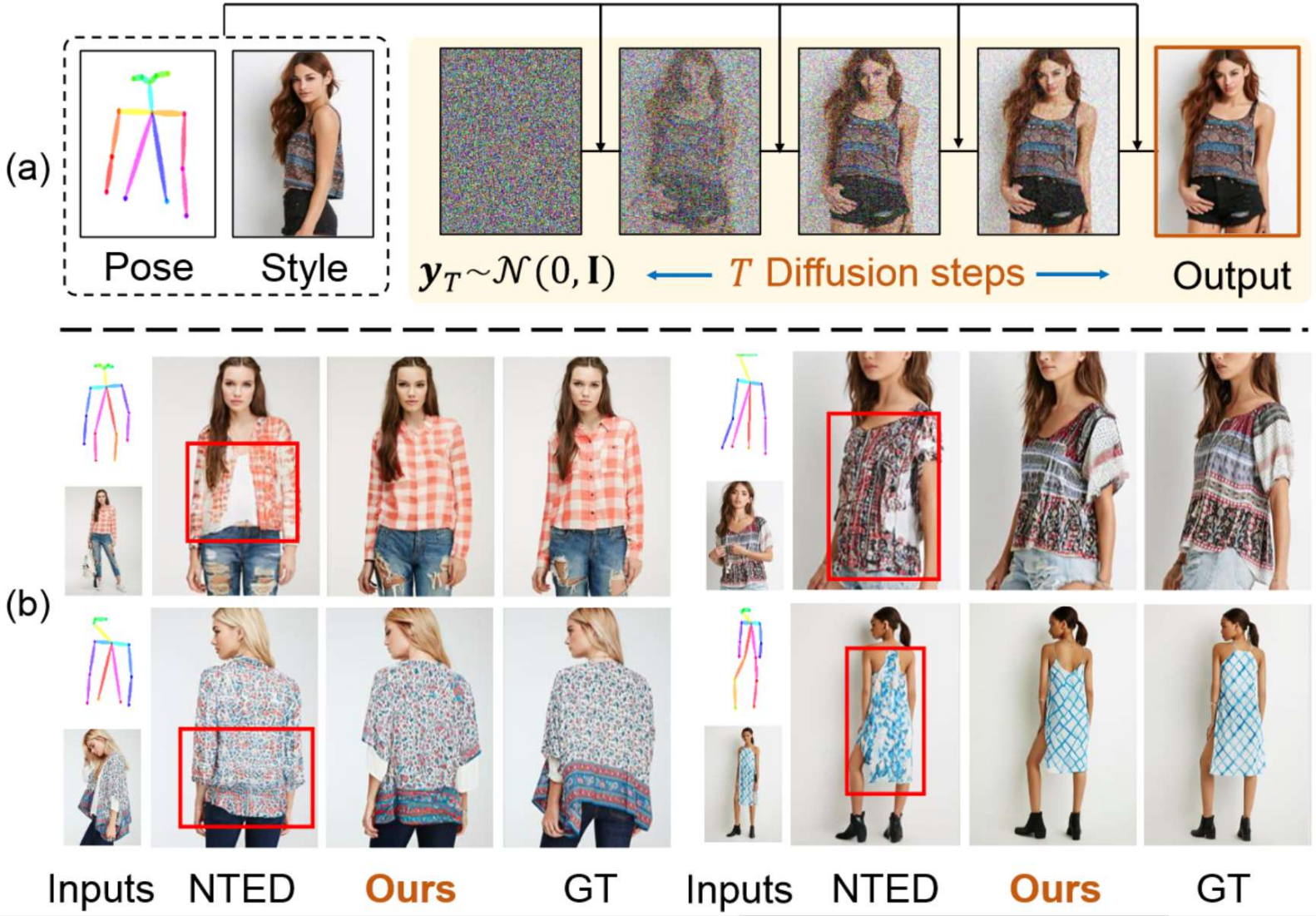
Person Image Synthesis via Denoising Diffusion Model

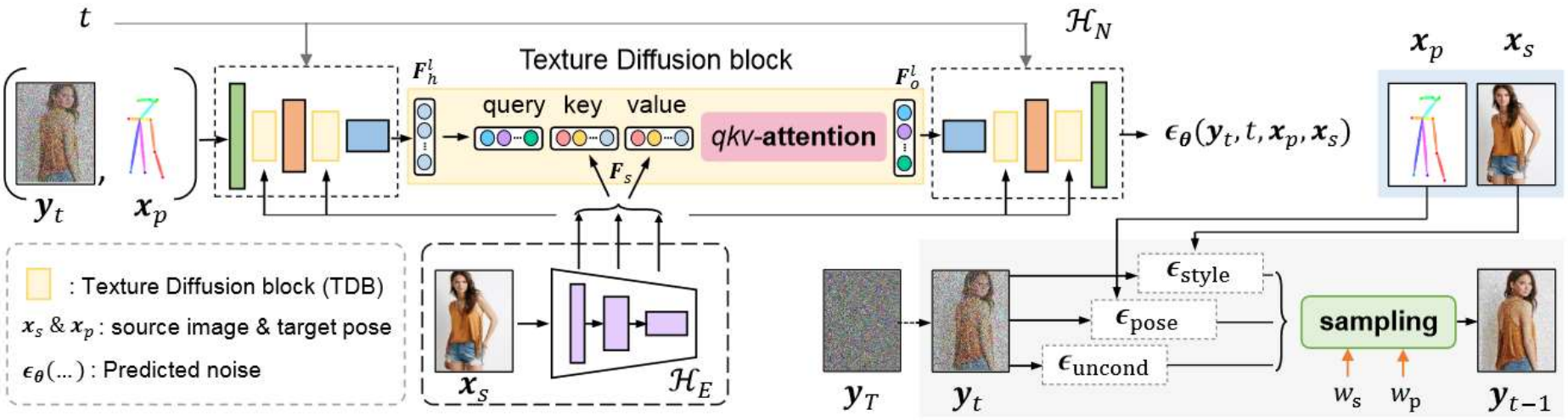
Ankan Kumar Bhunia[1] Salman Khan[1,2] Hisham Cholakkal[1] Rao Muhammad Anwer[1,4]
Jorma Laaksonen[4] Mubarak Shah[5] Fahad Shahbaz Khan[1,3]
[1]Mohamed bin Zayed University of AI, UAE [2]Australian National University, Australia
[3]Linköping University, Sweden [4]Aalto University, Finland [5]University of Central Florida, USA

汇报人：李雅超

Bhunia, Ankan Kumar, et al. "Person image synthesis via denoising diffusion model." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.

南京航空航天大學
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS



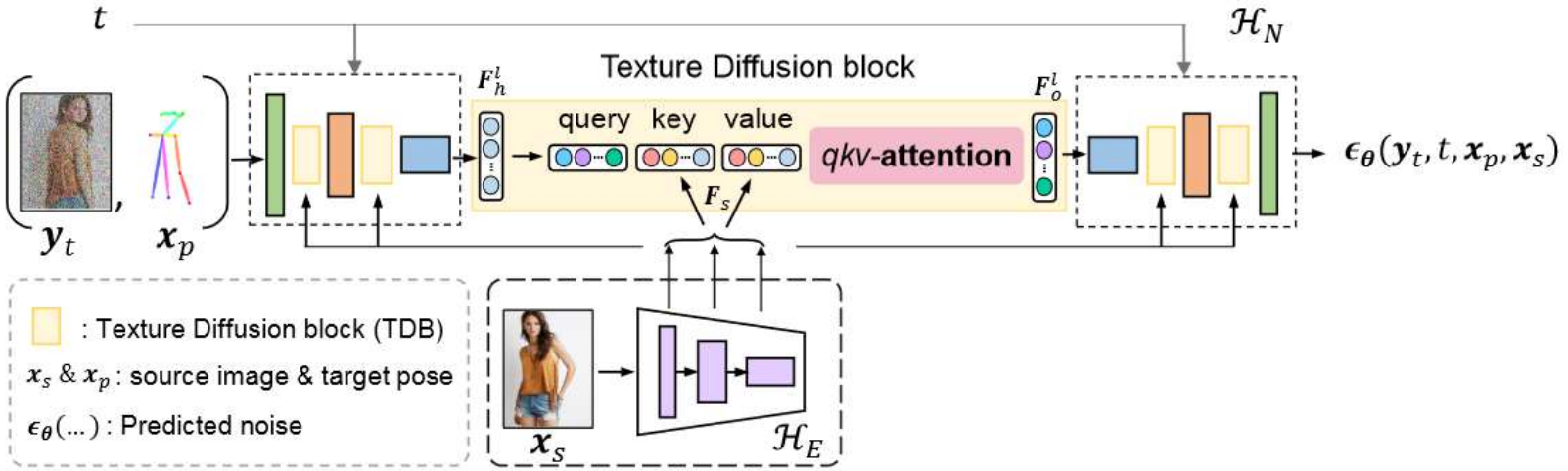


(a) Overview of our proposed PIDM architecture

(b) Disentangled Guidance based DDPM sampling

南京航空航天大學
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

Texture-Conditioned Diffusion Model



前向扩散分布:

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

逆向去噪分布:

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

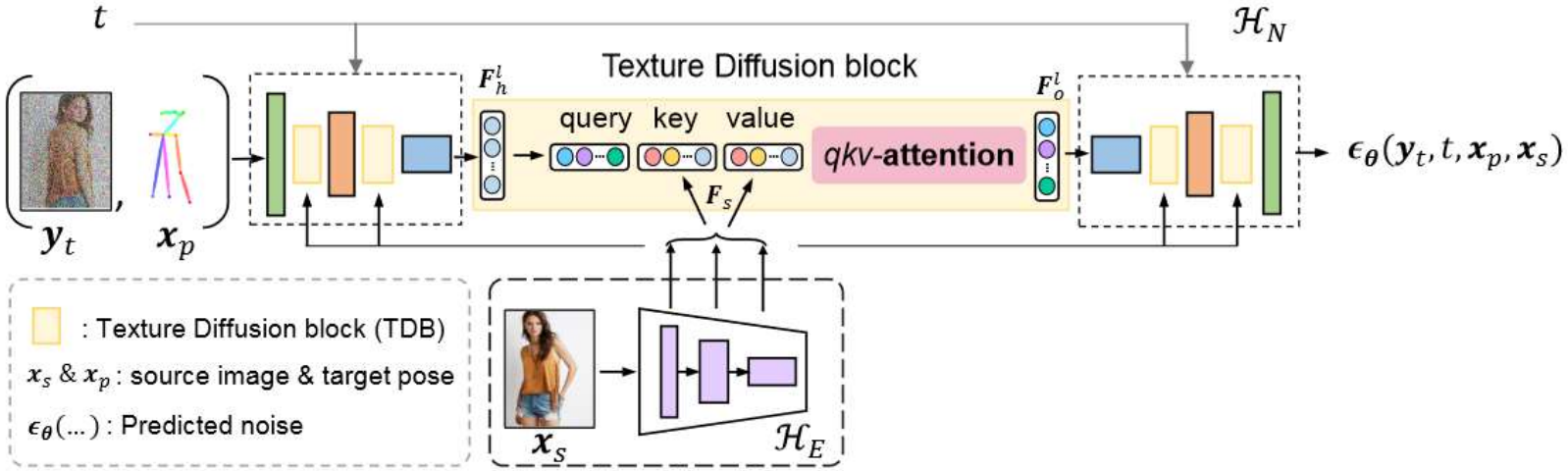
损失函数:

$$L_{\text{mse}} = \mathbb{E}_{t\sim[1,T],\boldsymbol{y}_0\sim q(\boldsymbol{y}_0),\epsilon} \|\epsilon - \epsilon_\theta(\boldsymbol{y}_t, t, \boldsymbol{x}_p, \boldsymbol{x}_s)\|^2. \quad (3)$$

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

南京航空航天大學
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

Texture Diffusion block

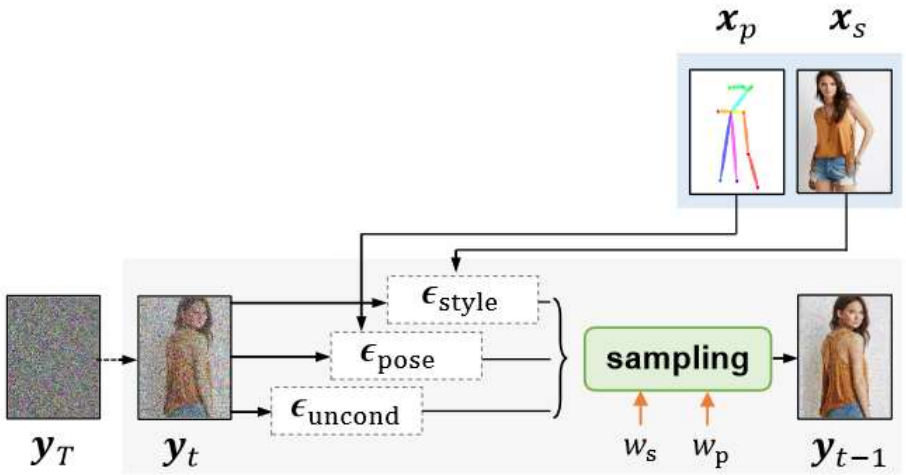


$$\boldsymbol{Q} = \phi_q^l(\boldsymbol{F}_h^l), \quad \boldsymbol{K} = \phi_k^l(\boldsymbol{F}_s), \quad \boldsymbol{V} = \phi_v^l(\boldsymbol{F}_s)$$

$$\boldsymbol{F}_{att}^l = \frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{C}}, \quad \boldsymbol{F}_{\boldsymbol{o}}^{\boldsymbol{l}} = \boldsymbol{W}^{\boldsymbol{l}}\mathrm{softmax}(\boldsymbol{F}_{att}^l)\boldsymbol{V} + \boldsymbol{F}_h^l, \tag{5}$$

南京航空航天大學
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

Disentangled Guidance based Sampling



$$\epsilon_{\text{uncond}} = \epsilon_\theta(\boldsymbol{y}_t, t, \emptyset, \emptyset)$$
$$\epsilon_{\text{pose}} = \epsilon_\theta(\boldsymbol{y}_t, t, \boldsymbol{x}_p, \emptyset) - \epsilon_{\text{uncond}}$$
$$\epsilon_{\text{style}} = \epsilon_\theta(\boldsymbol{y}_t, t, \emptyset, \boldsymbol{x}_s) - \epsilon_{\text{uncond}}$$

解耦的条件噪音:

$$\boldsymbol{\epsilon}_{\text{cond}} = \boldsymbol{\epsilon}_{\text{uncond}} + w_p \boldsymbol{\epsilon}_{\text{pose}} + w_s \boldsymbol{\epsilon}_{\text{style}}, \tag{6}$$

原始的无分类引导下的条件噪音:

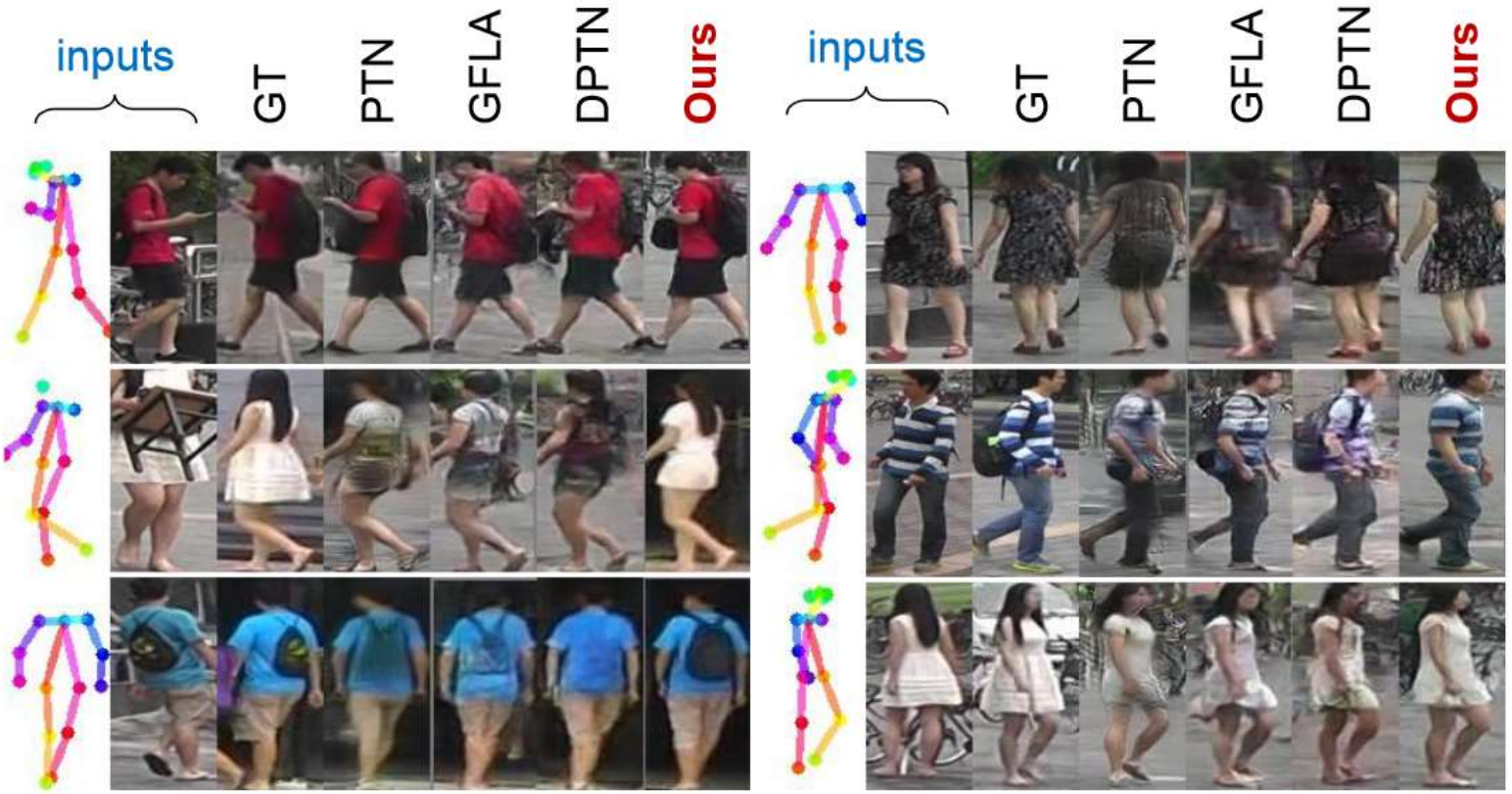
$$\begin{aligned}\bar{\boldsymbol{\epsilon}}_\theta(\mathbf{x}_t, t, y) &= \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, y) - \sqrt{1-\bar{\alpha}_t} w \nabla_{\mathbf{x}_t} \log p(y \mid \mathbf{x}_t) \\ &= \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, y) + w(\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, y) - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)) \\ &= (w+1)\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, y) - w\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)\end{aligned}$$

定量实验结果

Dataset	Methods	FID(↓)	SSIM(↑)	LPIPS(↓)
DeepFashion [11] (256 × 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
	PIDM (Ours)	**6.3671**	**0.7312**	**0.1678**
DeepFashion [11] (512 × 352)	CocosNet2 [29]	13.325	0.7236	0.2265
	NTED [18]	7.7821	0.7376	0.1980
	PIDM (Ours)	**5.8365**	**0.7419**	**0.1768**
Market-1501 [27] (128 × 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
	PIDM (Ours)	**14.451**	**0.3054**	**0.2415**

南京航空航天大學
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

可视化对比结果



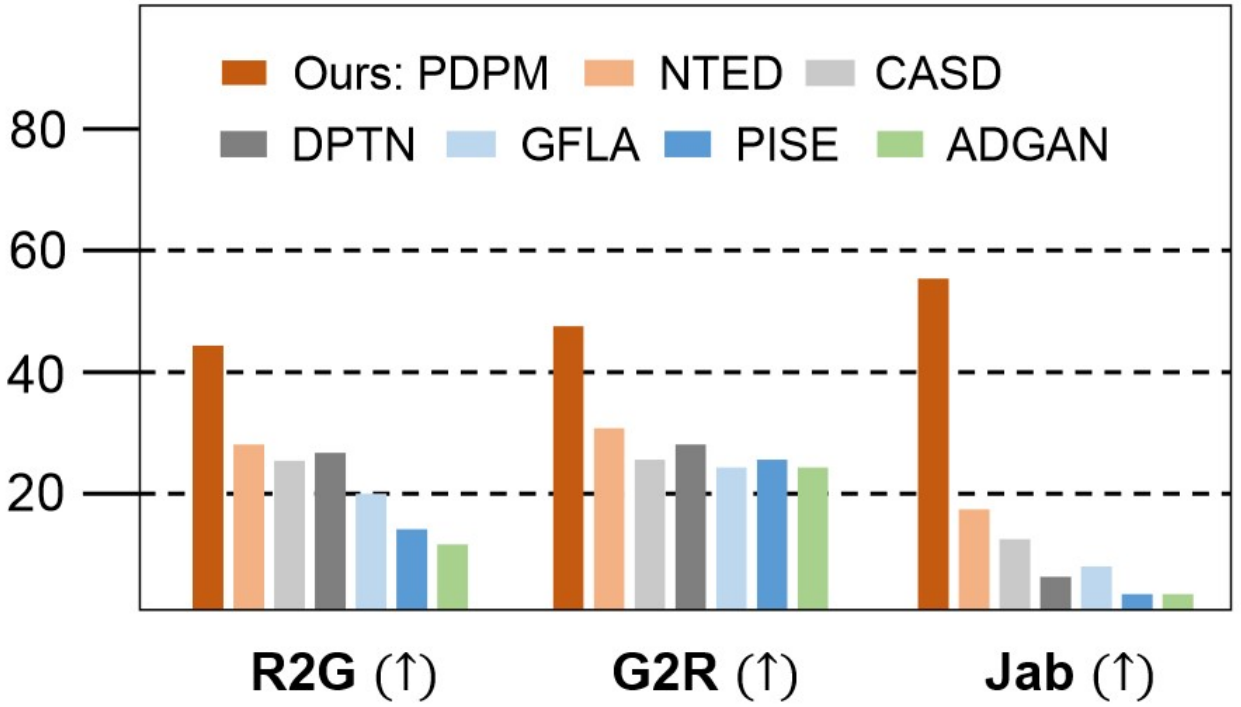

inputs
GT
ADGAN
PISE
GFLA
DPTN
CASD
NTED
Ours
inputs
GT
ADGAN
PISE
GFLA
DPTN
CASD
NTED
Ours
(1)
(2)
(3)
(4)
(5)
(6)
(7)
(8)
(9)
(10)

南京航空航天大學
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

User Study结果

30 generated images

30 real-world images



R2G: real-world to generate

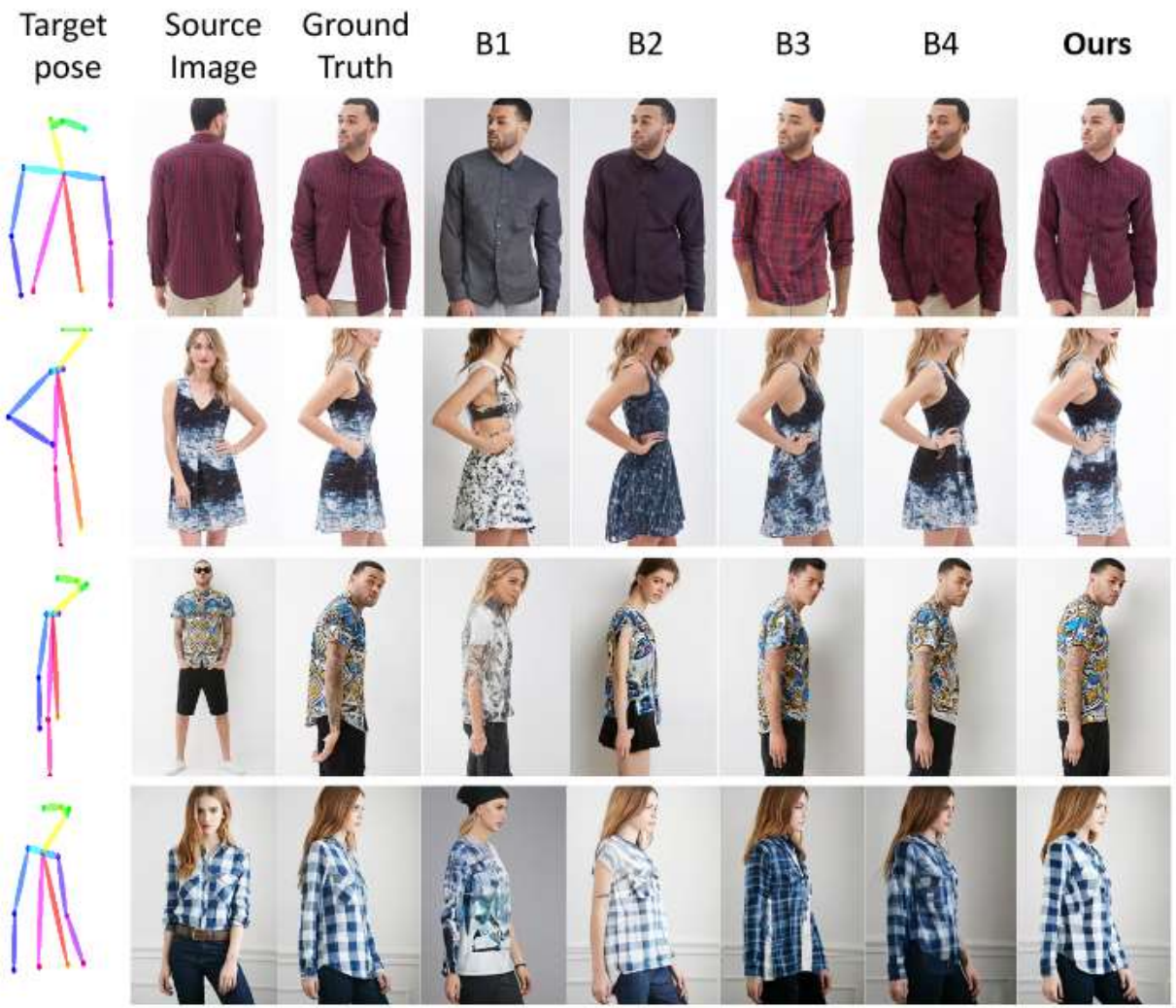
G2R: generate to real-world

Jab: the percentage of images considered the best among all models

消融实验

Methods	FID(↓)	SSIM(↑)	LPIPS(↓)
B1: `Baseline`† (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
Ours: Baseline + TDB + DCF-guidance	**6.3671**	**0.7312**	**0.1676**

消融实验



Appearance Control and Editing