NEURAL INFORMATION PROCESSING SYSTEMS

FashionR2R: Texture-preserving Rendered-to-Real Image Translation with Diffusion Models

Rui Hu[1*], Qian He[2*], Gaofeng He[2], Jiedong Zhuang[1],
Huang Chen[2], Huafeng Liu[1✉], Huamin Wang[2]
[1]Zhejiang University, [2]Style3D Research

NeurIPS 2024

Motivation

- Rendering methods
  - Modeling accuracy

  VS computation efficiency

From FaceSynthetics (ICCV 2021)

- Generative methods
  - Impressive authenticity
  - Poor controllability and editability

From LDM (CVPR 2022)

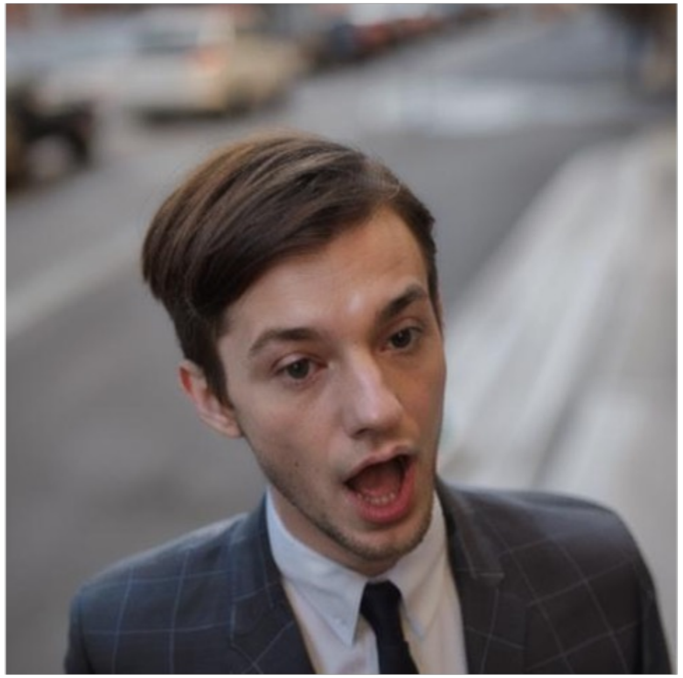
Problem Definition

- Rendered-to-Real image translation

Rendered

From FaceSynthetics (ICCV 2021)

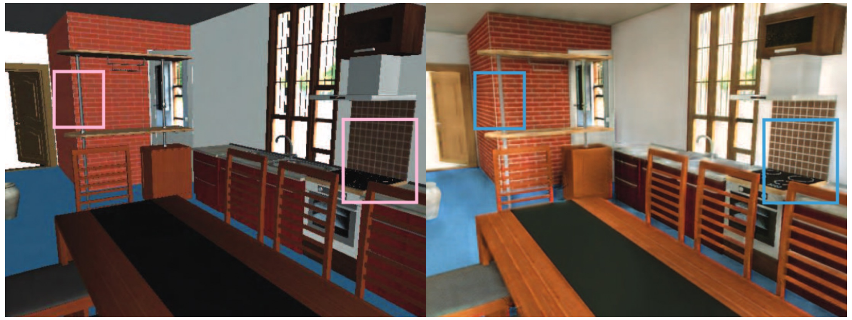
Real



Previous Works

- Rendered-to-real image translation

Deep CG2Real (ICCV 2019)



EPE (TPAMI 2022)

- General image-to-image translation & style transfer

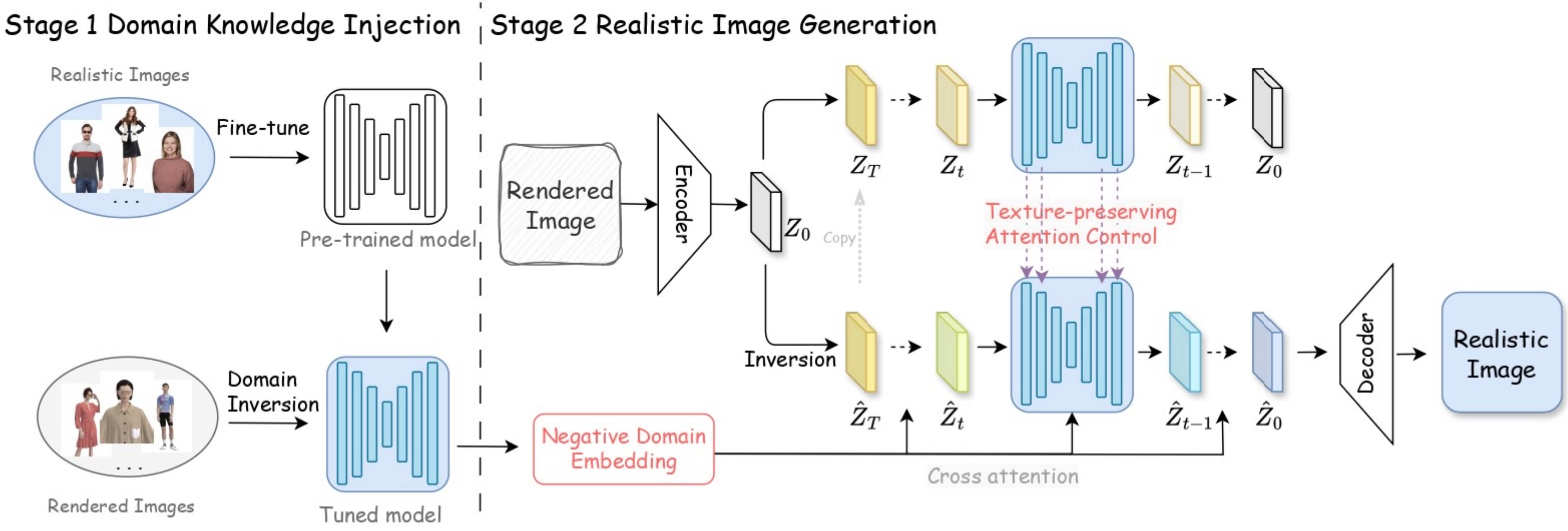
UNSB (ICLR 2024)

VCT (ICCV 2023)

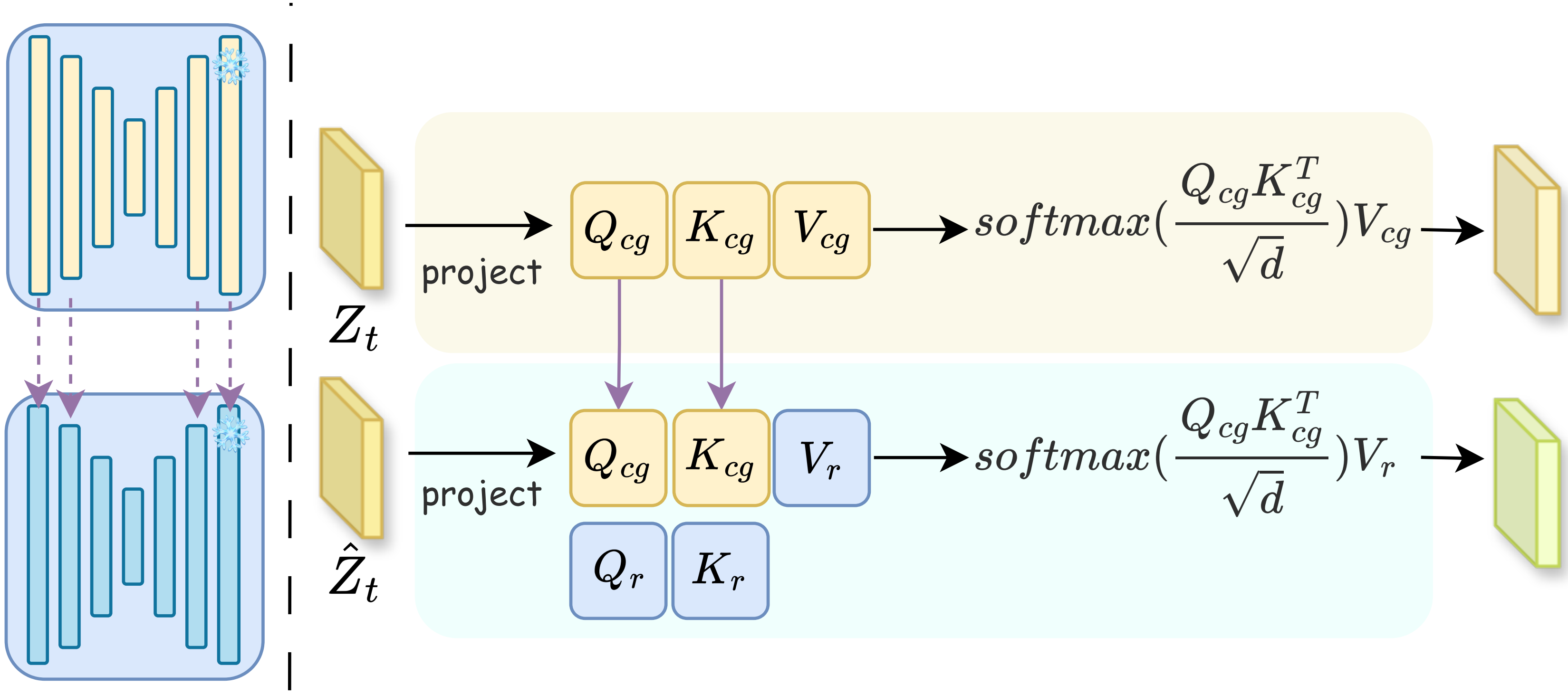
Main Idea

- Leverage generative prior from pretrained T2I diffusion models.
  - Adaptation to realistic image generation under the guidance of distilled rendered prior.
- Exploit decoupled features in the UNet structure.
  - A texture-preserving mechanism by extracting attention features from an inversion pipeline.

Framework


Stage 1 Domain Knowledge Injection
Stage 2 Realistic Image Generation
Realistic Images
Fine-tune
Pre-trained model
Rendered Images
Domain Inversion
Tuned model
Rendered Image
Encoder
Z_0
Copy
Z_T
Z_t
Z_{t-1}
Z_0
Texture-preserving Attention Control
Inversion
$\hat{Z}_T$
$\hat{Z}_t$
$\hat{Z}_{t-1}$
$\hat{Z}_0$
Decoder
Realistic Image
Negative Domain Embedding
Cross attention

Texture-preserving Attention Control



$$\widehat{Q^t}, \widehat{K^t} = TAC\left(Q_{cg}^t, K_{cg}^t, Q_r^t, K_r^t, t\right) = \begin{cases} Q_{cg}^t, K_{cg}^t & \text{if } t < \gamma T, f > F \\ Q_r^t, K_r^t & \text{otherwise} \end{cases}$$

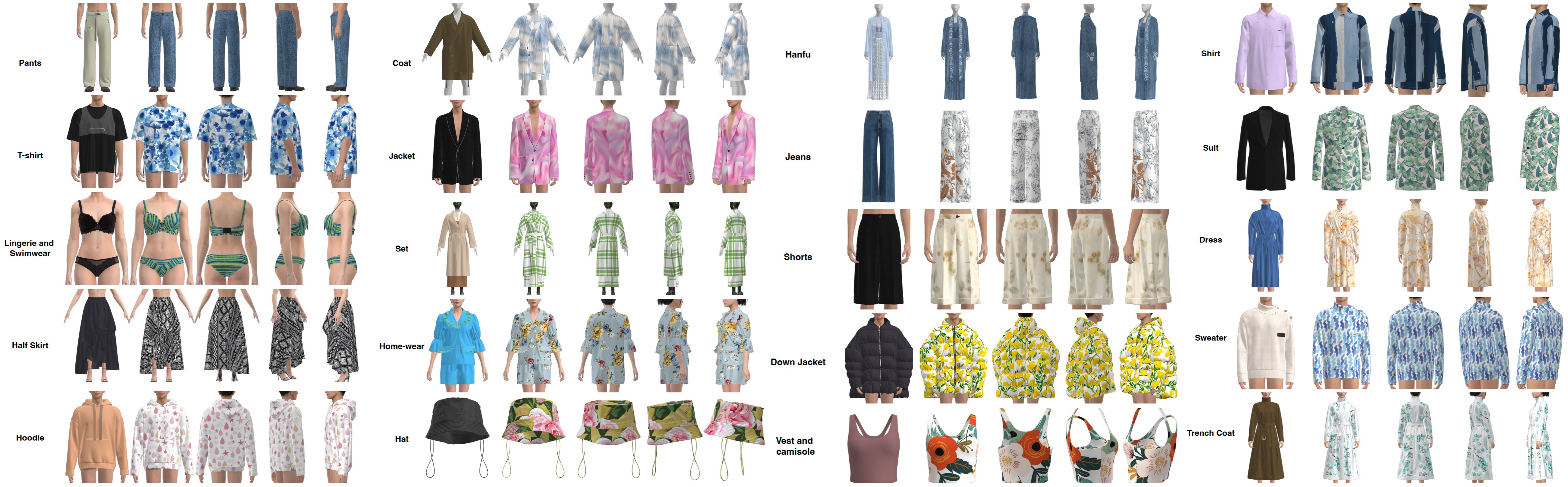
Evaluation

- Datasets
  - Face Synthetics & SynFashion.

- Metrics
  - KID (to real), LPIPS & SSIM (to rendered).
- User studies
  - Overall realism, image quality and consistency.

SynFashion Dataset

- Collected via Style3D Studio
  - 10k rendered images, 20 categories, 375 3D projects, 500 textures



Qualitative Results – Face Synthetics

Qualitative Results – SynFashion

Quantitative Results

- Quantitative comparisons

Dataset	Face Synthetics			SynFashion		
	KID↓(std)	LPIPS↓(std)	SSIM↑(std)	KID↓(std)	LPIPS↓(std)	SSIM↑(std)
CUT [10]	80.553 (2.447)	0.365 (0.073)	0.664 (0.079)	<u>59.238</u> (1.599)	0.170 (0.060)	0.847 (0.067)
SANTA [11]	90.390 (2.929)	0.387 (0.079)	0.618 (0.104)	61.636 (1.628)	0.294 (0.067)	0.741 (0.082)
VCT [13]	<u>74.445</u> (2.273)	**0.096** (0.027)	0.807 (0.072)	59.489 (1.499)	0.178 (0.058)	0.807 (0.085)
UNSB [12]	76.389 (2.465)	0.229 (0.069)	<u>0.818</u> (0.070)	59.496 (1.453)	<u>0.130</u> (0.040)	**0.891** (0.054)
Ours	**73.871** (1.973)	<u>0.121</u> (0.035)	**0.831** (0.068)	**54.720** (1.362)	**0.067** (0.025)	<u>0.881</u> (0.055)

- User studies - percentage of votes each method are preferred to ours

Dataset	Face Synthetics			SynFashion		
	Overall Realism	Image Quality	Consistency	Overall Realism	Image Quality	Consistency
CUT	0.529%	0.529%	13.175%	8.994%	6.878%	16.931%
SANTA	0.922%	1.383%	12.304%	3.333%	5.238%	11.571%
VCT	5.952%	14.286%	20.714%	2.041%	6.122%	18.367%
UNSB	4.511%	6.767%	21.278%	9.821%	9.821%	26.607%

Ablation Studies

- Visual examples

- Numerical results

Dataset	Face Synthetics			SynFashion		
	KID↓(std)	LPIPS↓(std)	SSIM↑(std)	KID↓(std)	LPIPS↓(std)	SSIM↑(std)
w/o source DKI	77.376 (2.063)	0.107 (0.029)	0.857 (0.059)	58.520 (1.902)	0.059 (0.019)	0.903 (0.065)
w/o target DKI	78.927 (2.134)	0.114 (0.031)	0.845 (0.063)	60.186 (1.623)	0.064 (0.022)	0.897 (0.056)
w/o TAC	69.349 (1.485)	0.253 (0.070)	0.720 (0.085)	51.392 (1.083)	0.183 (0.047)	0.794 (0.074)
Ours	73.831 (1.973)	0.121 (0.035)	0.831 (0.068)	54.720 (1.362)	0.067 (0.025)	0.881 (0.055)

Conclusion

- A novel framework for rendered-to-real fashion image translation.
  - Generative prior from pretrained diffusion models.
- Two-stage: Domain Knowledge Injection and Realistic Image Generation.
  - DKI: positive domain finetuning & negative domain embedding.
  - RIG: texture-preserving attention control.
- A high-quality rendered fashion image dataset: SynFashion.