Pose guided human image synthesis by view disentanglement and enhanced weighting loss

Mohamed Ilyes Lakhal[1], Oswald Lanz[2], Andrea Cavallaro[1]

Published in: *Workshop on Human Behaviour Understanding, in conjunction with ECCV 2018*

Centre for Intelligent Sensing

Queen Mary University of London

Queen Mary University of London[1]
Fondazione Bruno Kessler[2]

CIS centre for intelligent sensing

Queen Mary
University of London

Pose guided human image synthesis

Source Image

How to generate?

Target Pose
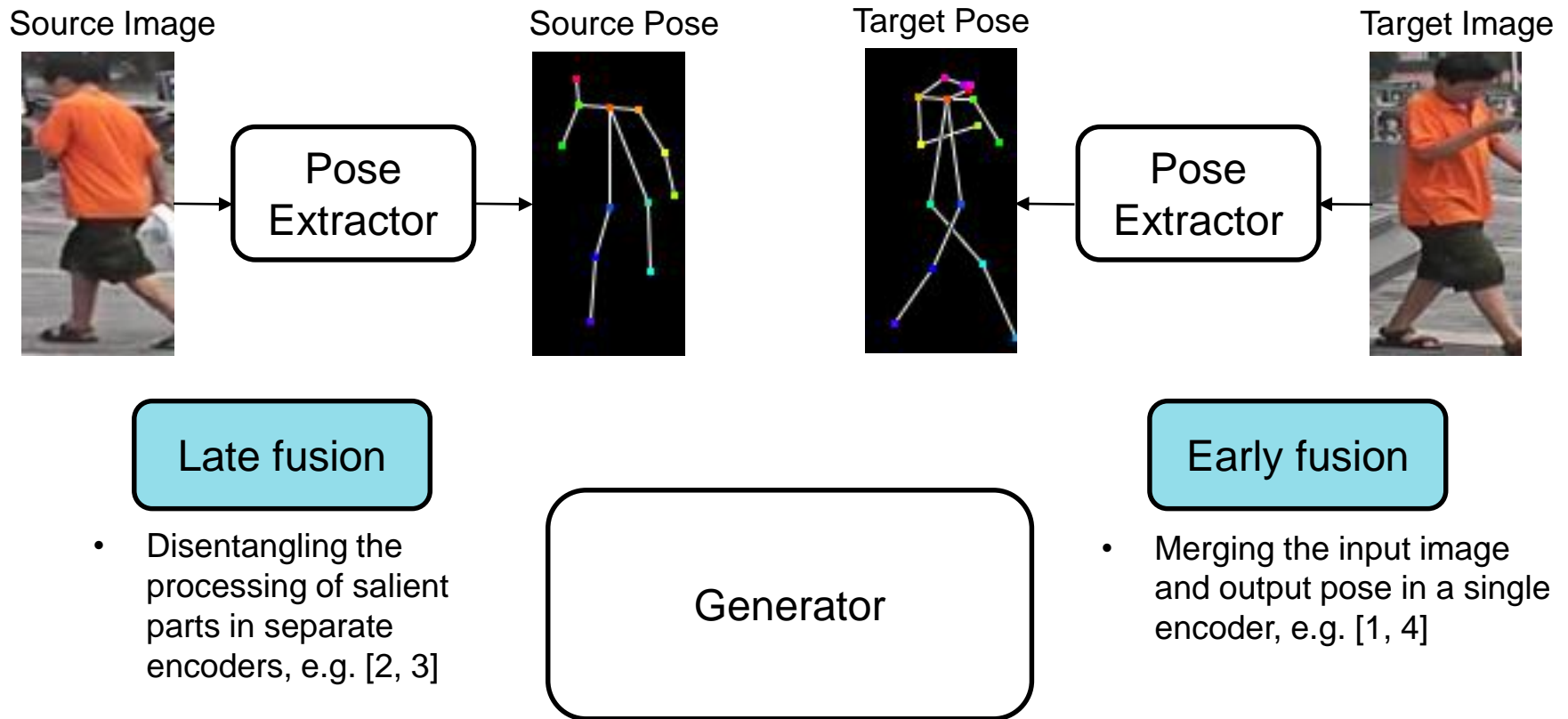
$\oplus$

Generated Image

Target Image

$\approx$

Related work

Source Image

Source Pose

Target Pose

Target Image

Pose Extractor

Pose Extractor

Late fusion

- Disentangling the processing of salient parts in separate encoders, e.g. [2, 3]

Generator

Early fusion

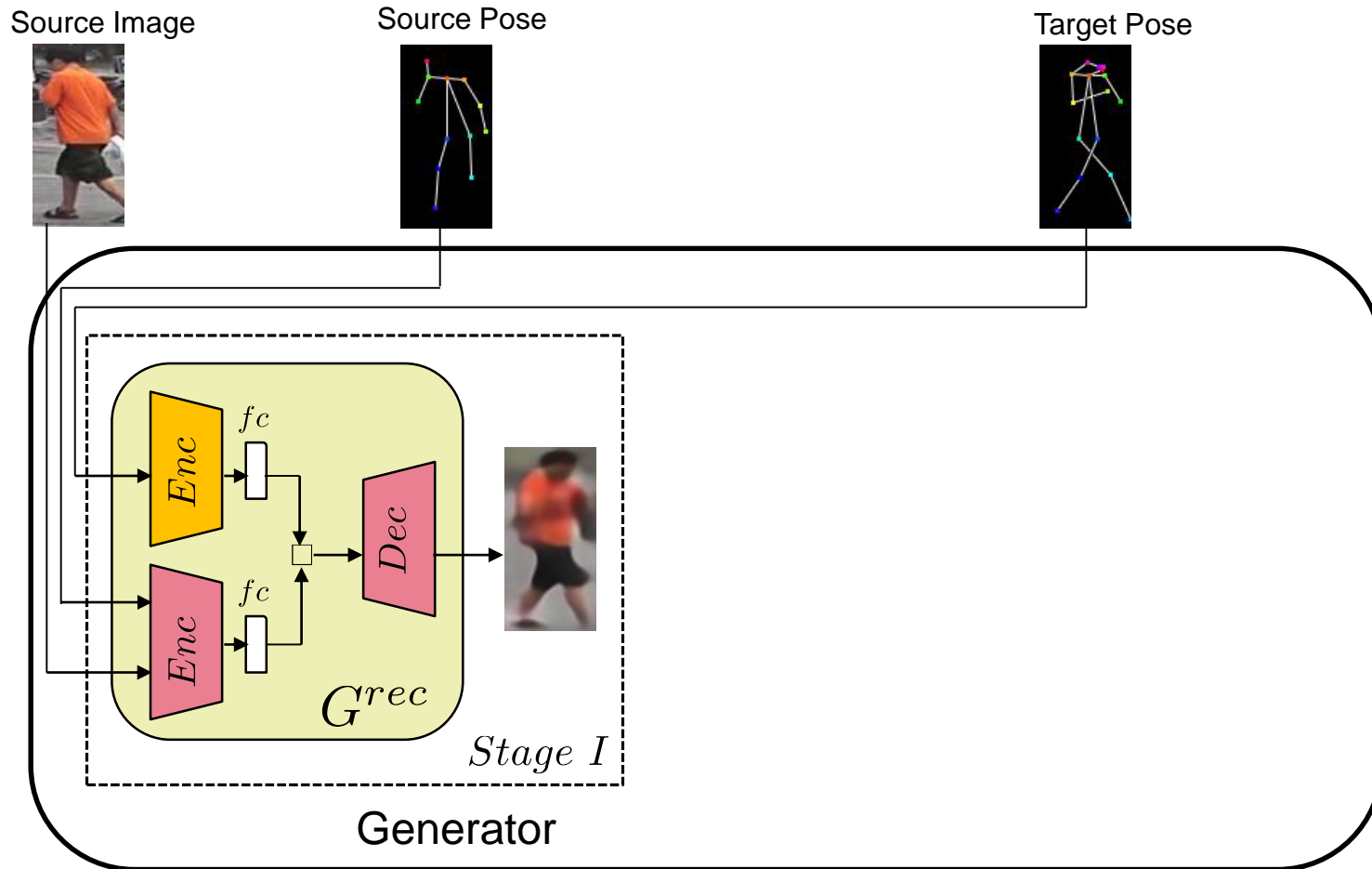- Merging the input image and output pose in a single encoder, e.g. [1, 4]

[1] L. Ma *et al*. Pose guided person image generation. In NIPS, 2017.
[2] A. Siarohin *et al*. Deformable GANs for Pose-based Human Image Generation. In CVPR 2018.
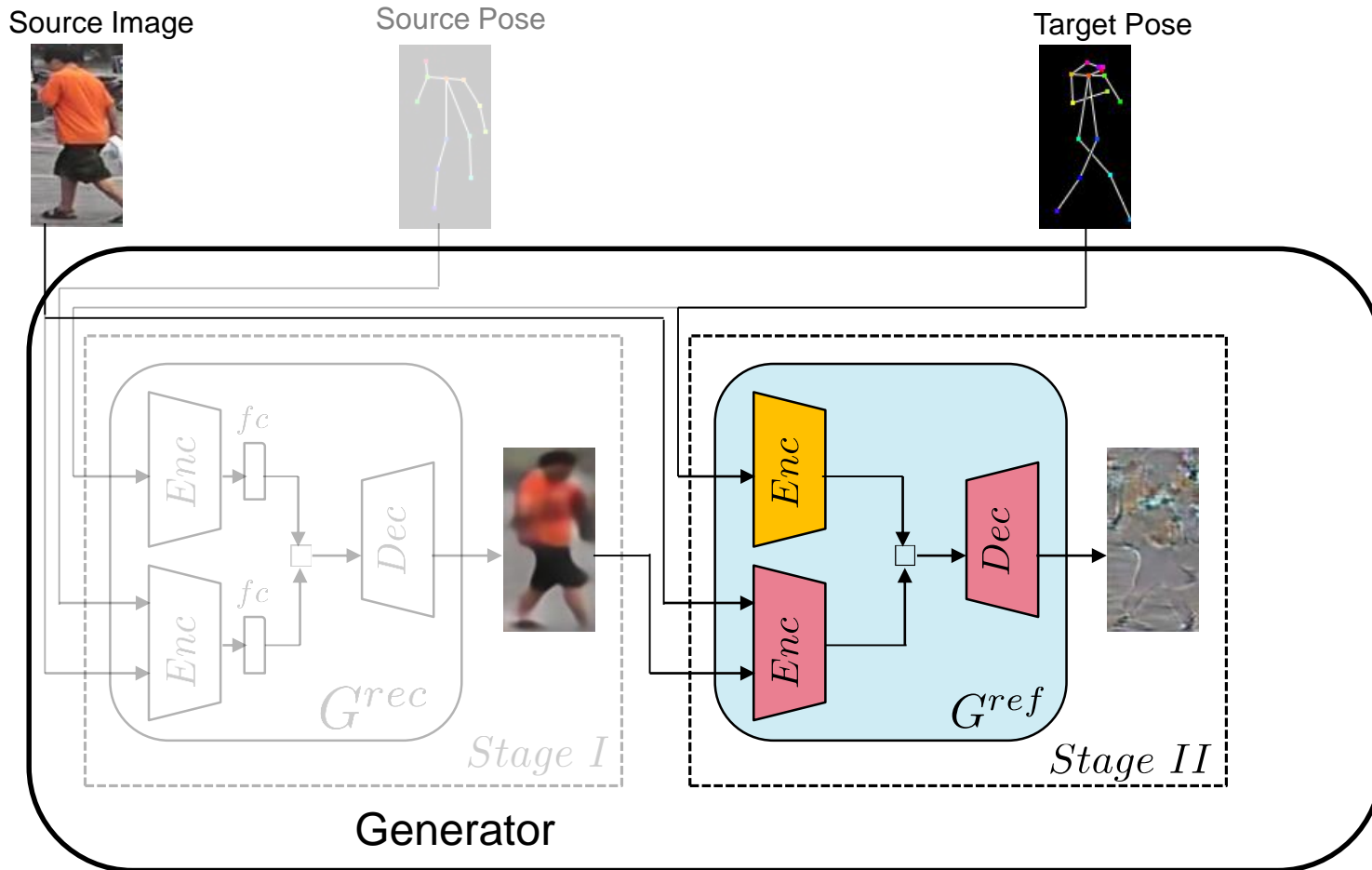[3] L. Ma *et al*. Disentangled Person Image Generation. In CVPR 2018.
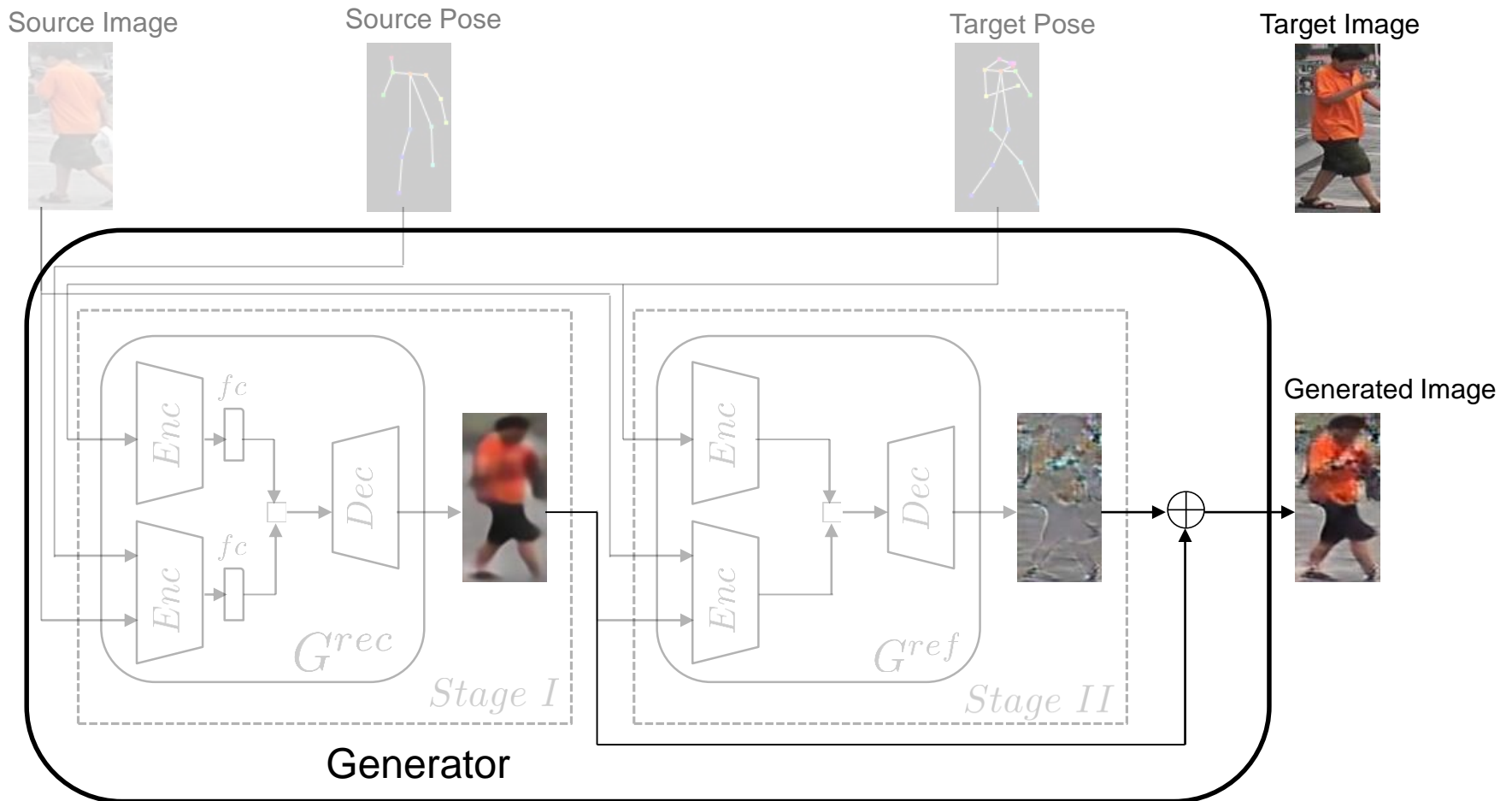[4] B. Zhao *et al*. Multi-View Image Generation from a Single-View. In Arxiv 2017.

CIS centre for intelligent sensing

Queen Mary
University of London

Proposed VDG model – **Stage I**

Source Image

Source Pose

Target Pose

fc

Enc

fc

Enc

Dec

G^{rec}

$Stage\ I$

Generator

CIS centre for intelligent sensing

Queen Mary
University of London

Proposed VDG model – **Stage II**

Source Image

Source Pose

Target Pose

fc

Enc

fc

Enc

Dec

G^{rec}

Stage I

Enc

Enc

Dec

G^{ref}

Stage II

Generator

Proposed VDG model – **Final prediction**

Source Image

Source Pose

Target Pose

Target Image

Generated Image

fc

fc

Enc

Enc

Dec

G^{rec}

Stage I

Enc

Enc

Dec

G^{ref}

Stage II

Generator

CIS centre for intelligent sensing

Queen Mary
University of London

Training losses

- Stage I:

$$\mathcal{L}_{G^{rec}} = ||\hat{I}_b^R - I_b||_1$$

- Stage II:

$$\mathcal{L}_{G^{ref}} = \mathcal{L}_{adv} + \lambda \mathcal{L}_{img}$$

$$\mathcal{L}_{img} = \alpha \mathcal{L}_{SSIM}^{mask} + (1 - \alpha) \mathcal{L}_1^{mask}$$

$$\mathcal{L}_1^{mask} = ||(\hat{I}_b^G - I_b) \odot (1 + \mathbb{M}_b)||_1$$

$$\mathcal{L}_{SSIM}^{mask} = \mathcal{L}_{SSIM}(\hat{I}_b \odot (1 + \mathbb{M}_b), I_b \odot (1 + \mathbb{M}_b))$$

Experiments – Settings

- Datasets: Market-1501 and DeepFashion

- Evaluation measures:
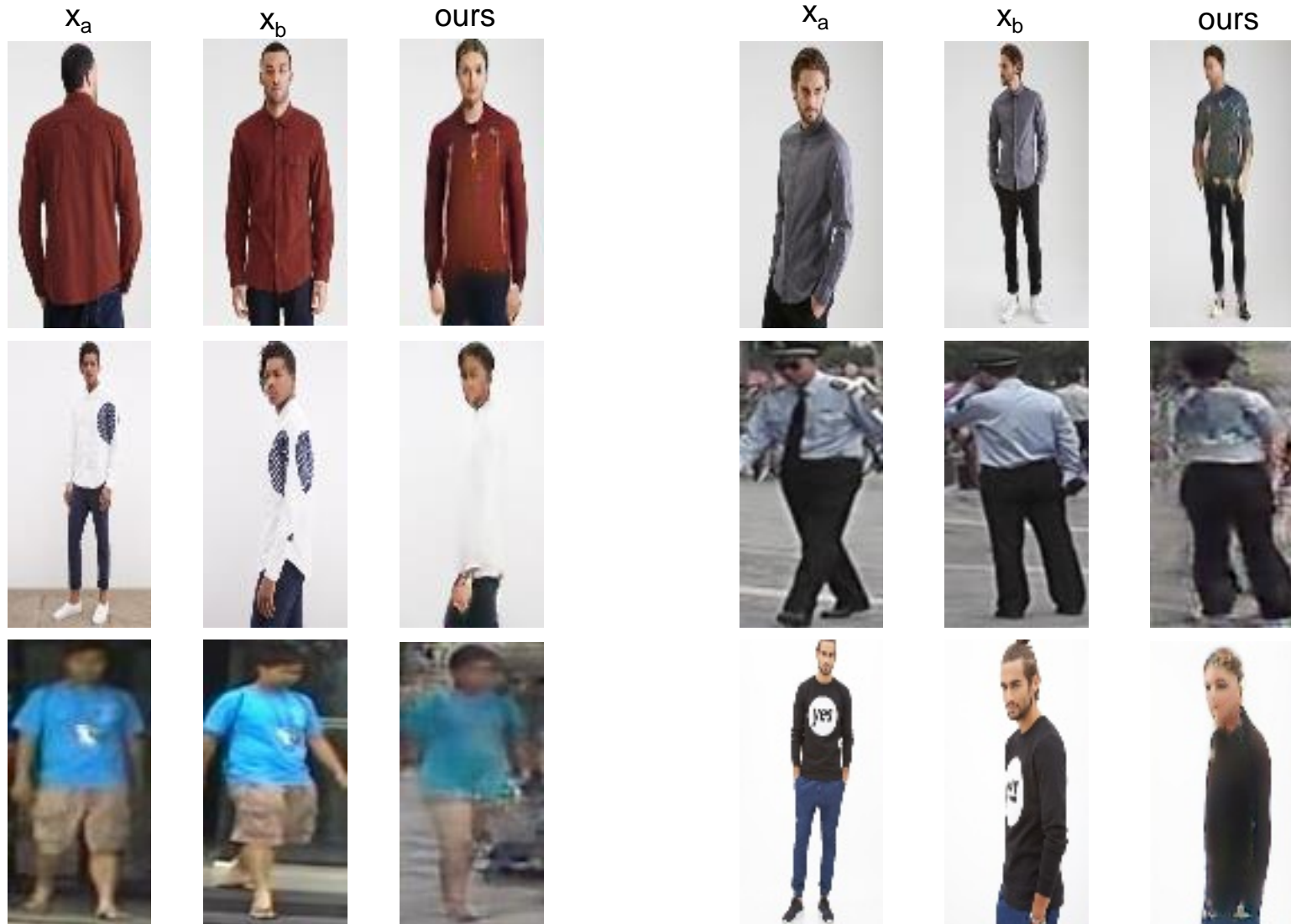 - SSIM score
  - Inception score

Experiments – Quantitative

Method	Market-1501				DeepFashion	
	SSIM	IS	mask-SSIM	mask-IS	SSIM	IS
PG^2 [1]	.252	**4.015**	.771	**3.555**	.641	3.187
Def-GAN [2]	**.290**	2.990	**.798**	3.544	.665	3.420
PDIG [3]	.099	3.483	.614	3.491	.614	3.228
VDG^{L1}	.224	3.733	.767	3.503	.700	3.428
VDG$^{mask-L1}$	.238	3.933	.768	3.542	.690	3.429
VDG	.238	4.007	.775	3.354	**.708**	3.003
VDG$_w$	.266	3.453	.783	3.227	.702	**3.491**

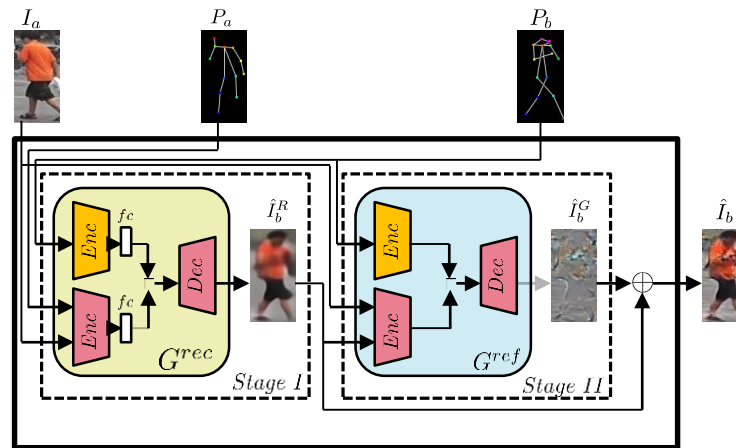[1] L. Ma *et al*. Pose guided person image generation. In NIPS, 2017.
[2] A. Siarohin *et al*. Deformable GANs for Pose-based Human Image Generation. In CVPR 2018.
[3] L. Ma *et al*. Disentangled Person Image Generation. In CVPR 2018.

CIS centre for intelligent sensing

Queen Mary
University of London

Experiments – Qualitative Results

x_a $\qquad$ x_b $\qquad$ ours $\qquad\qquad$ x_a $\qquad$ x_b $\qquad$ ours

Conclusion

- Proposed a late fusion generator to explicitly separate the processing of the input and the target in the encoder.
- Presented a new loss function.

Thank you!