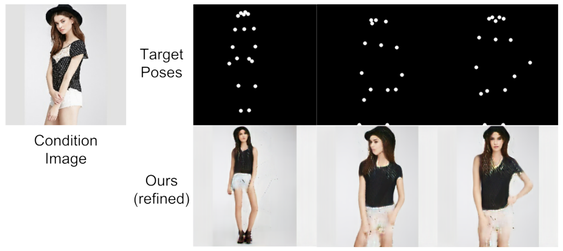
Pose Guided Person Image Generation

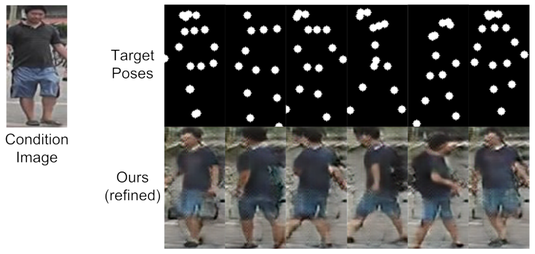
Liqian Ma[1], Xu Jia[2*], Qianru Sun[3*], Bernt Schiele[3], Tinne Tuytelaars[2], Luc Van Gool[1,4]
[1]KU-Leuven/PSI, TRACE (Toyota Res in Europe) [2]KU-Leuven/PSI, IMEC
[3]Max Planck Institute for Informatics, Saarland Informatics Campus [4]ETH Zurich

Introduction

- **Task:** Synthesize person images in arbitrary poses, based on an image of that person and a novel pose.
- **Motivation:** Provide users more control over the generation process.
- **Key idea:** Guide the generation process explicitly by an appropriate representation of that intention.



On DeepFashion dataset



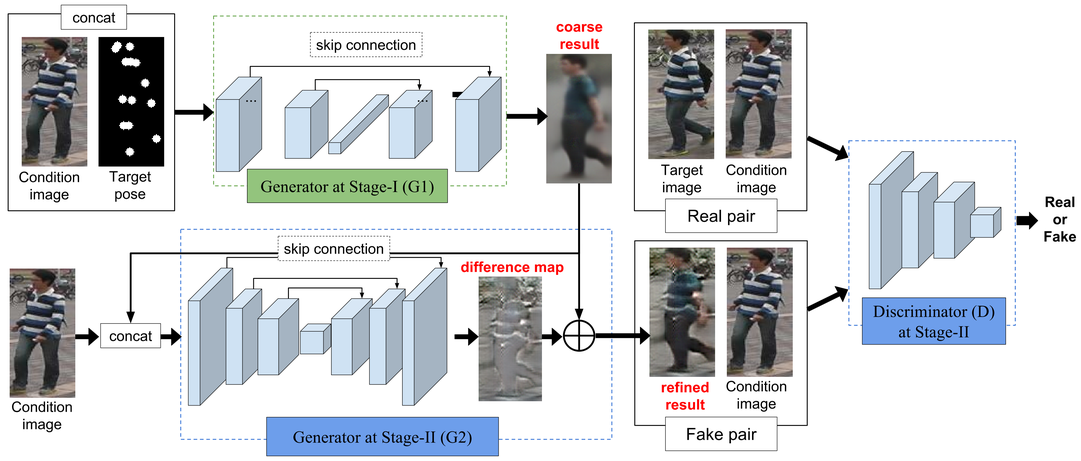
On Market-1501 dataset

- **Contributions**
  - i) A novel task of conditioning image generation on a reference image and an intended pose.
  - ii) A novel mask loss is proposed to encourage the model to focus on transferring the human body appearance instead of background information.
  - iii) Divide the problem into two stages, with stage-I focusing on global structure and stage-II on filling in appearance details.

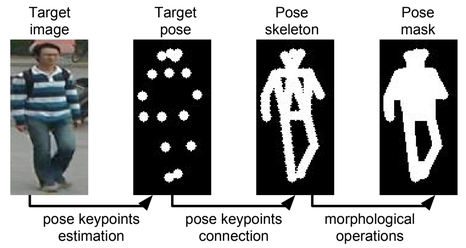
- **Dataset**
  - i) DeepFashion[1]: in-shop clothes retrieval dataset with 256x256 resolution.
  - ii) Market-1501[2]: person re-identification dataset with 128x64 resolution.

Two-stage framework

- **Pose Guided Person Generation Network (PG²)**



- **Optimization loss**



  - i) Stage-I loss

$$\mathcal{L}_{\mathrm{G1}} = \|(\mathrm{G1}(I_A, P_B) - I_B) \odot (1 + M_B)\|_1$$

  - ii) Stage-II loss

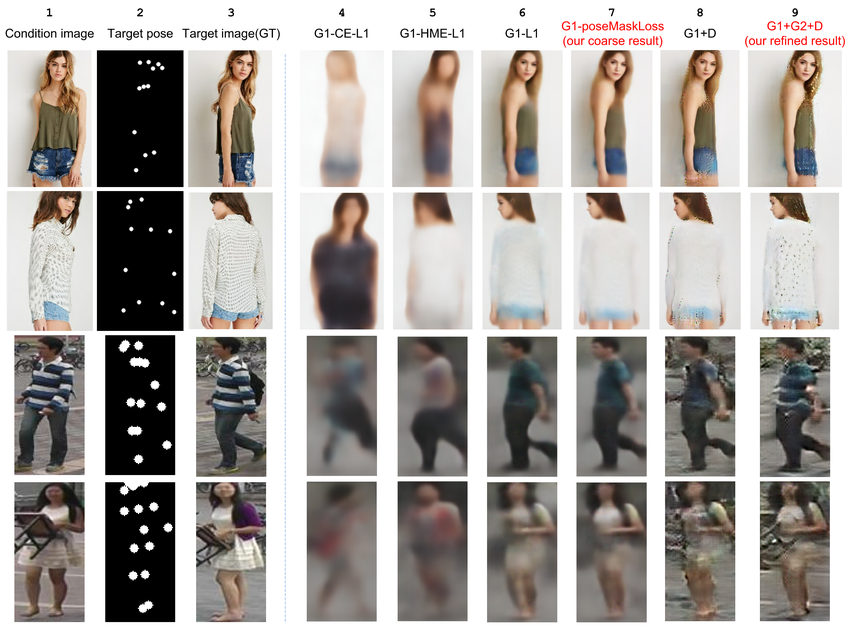
$$\mathcal{L}^{\mathrm{D}}_{adv} = \mathcal{L}_{bce}(\mathrm{D}(I_A, I_B), 1) + \mathcal{L}_{bce}(\mathrm{D}(I_A, \mathrm{G2}(I_A, \hat{I}_{B1})), 0)$$

$$\mathcal{L}^{\mathrm{G}}_{adv} = \mathcal{L}_{bce}(\mathrm{D}(I_A, \mathrm{G2}(I_A, \hat{I}_{B1})), 1)$$

$$\mathcal{L}_{\mathrm{G2}} = \mathcal{L}^{\mathrm{G}}_{adv} + \lambda\|(\mathrm{G2}(I_A, \hat{I}_{B1}) - I_B) \odot (1 + M_B)\|_1$$

Generation results

- **Qualitative results**



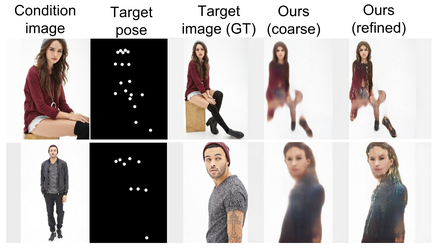
- **Quantitative results**

Table 1: Quantitative evaluation. For all measures, higher is better.

	DeepFashion		Market-1501			
Model	SSIM	IS	SSIM	IS	mask-SSIM	mask-IS
G1-CE-L1	0.694	2.395	0.219	2.568	0.771	2.455
G1-HME-L1	0.735	2.427	0.294	3.171	0.802	2.508
G1-L1	0.735	2.427	0.304	3.006	0.809	2.455
G1-poseMaskLoss	0.779	2.668	0.340	3.326	0.817	2.682
G1+D	0.761	3.091	0.283	3.490	0.803	3.310
G1+G2+D	0.762	3.090	0.253	3.460	0.792	3.435

- **Further analysis**

Comparison examples with [3]



Our failure cases on DeepFashion

[1] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In CVPR, pages 1096–1104, 2016.
[2] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In ICCV, pages 1116–1124, 2015.
[3] Bo Zhao, Xiao Wu, Zhi-Qi Cheng, Hao Liu, and Jiashi Feng. Multi-view image generation from a single-view. arXiv, 1704.04886, 2017.