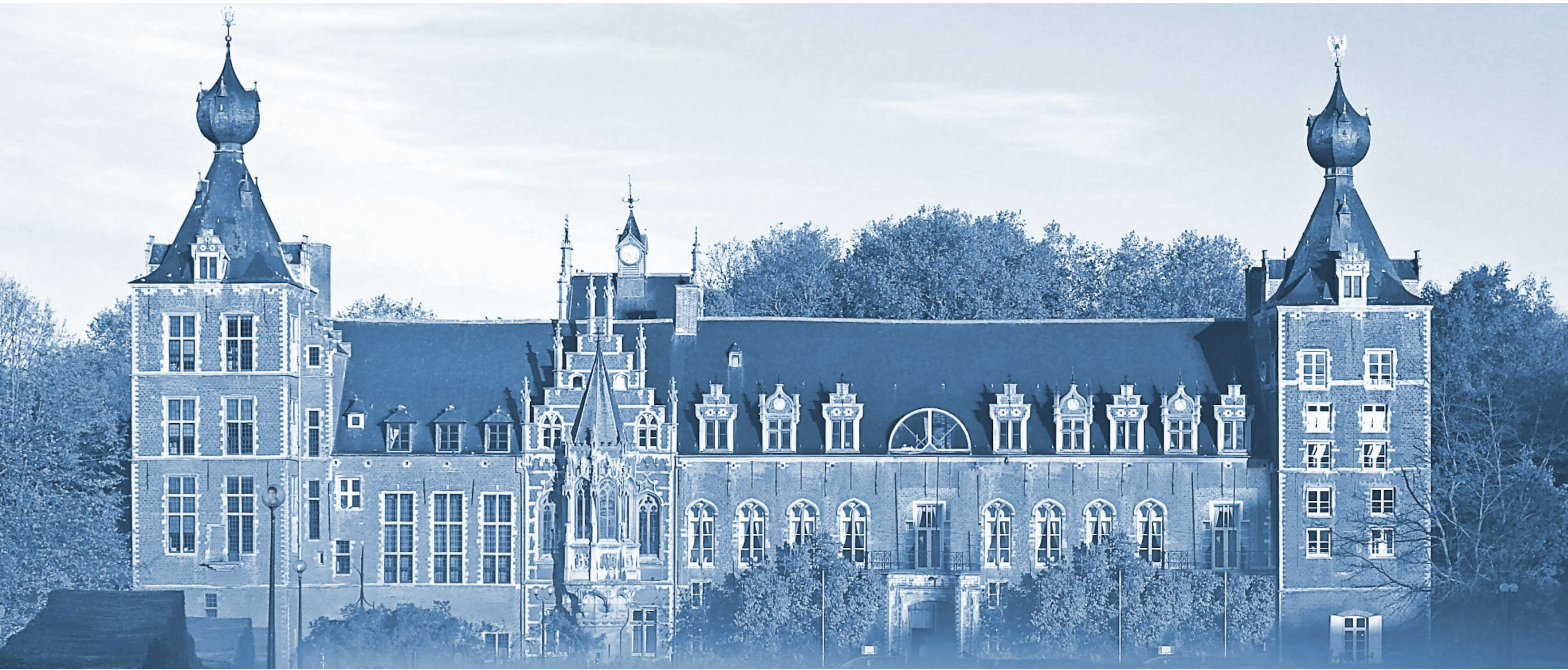
Pose Guided Person Image Generation

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

Accepted at NIPS 2017

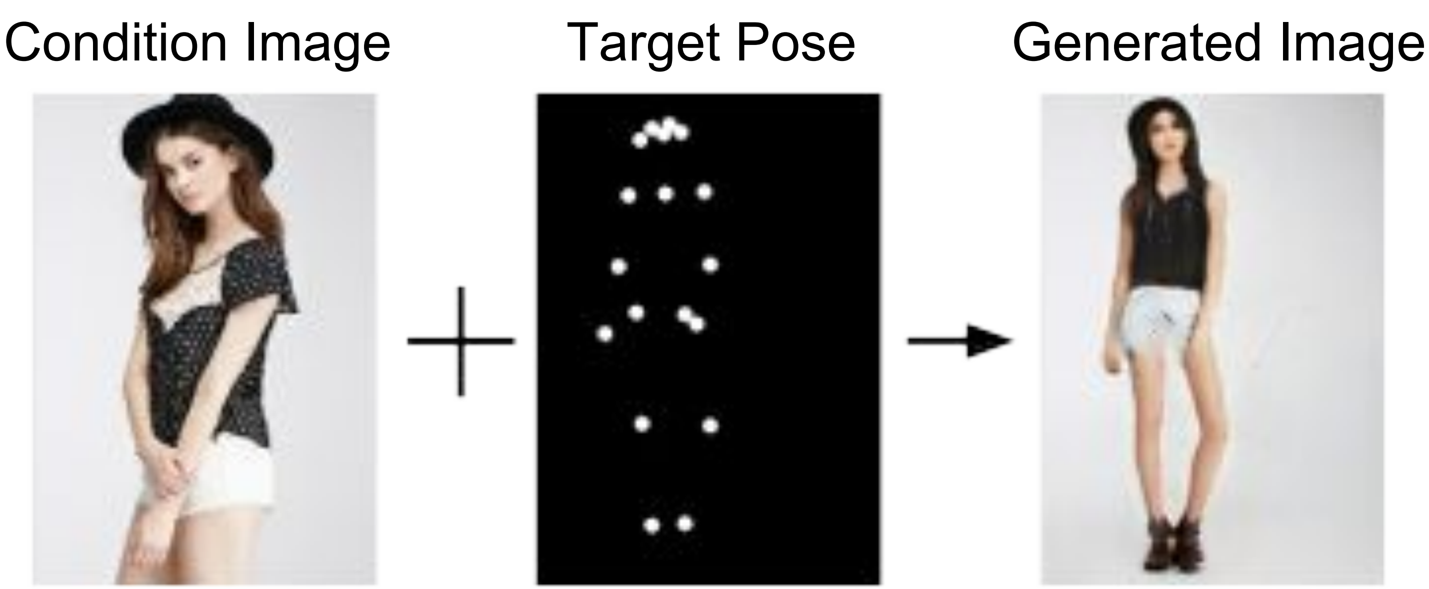


Outline

- Problem Statement.
- Related Work.
- Method.
- Experiments.
- Conclusions.

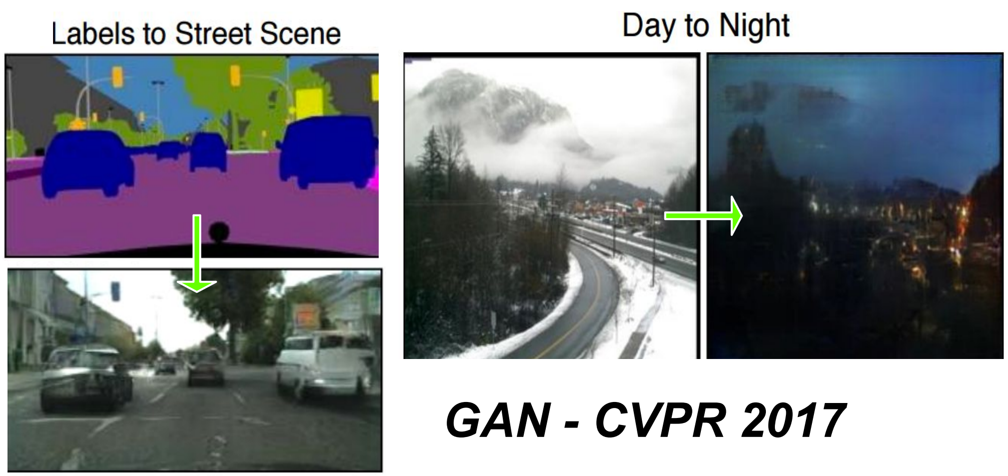
Problem Statement

- **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.



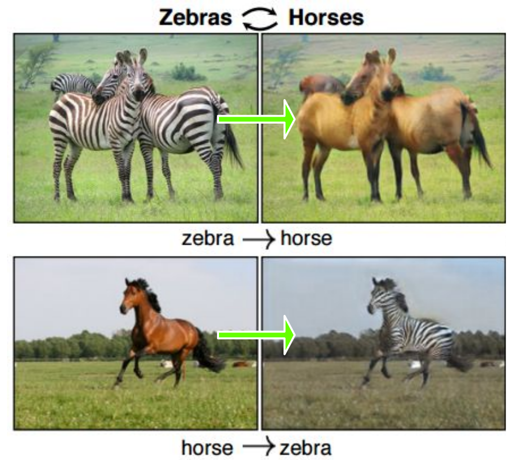
Related Work

- **Image --> Image**



GAN - CVPR 2017

CVPR 2017, Image-to-Image Translation with Conditional Adversarial Networks



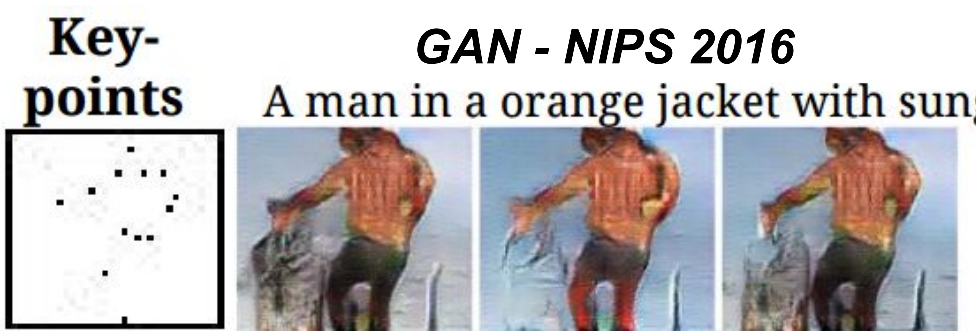
CycleGAN - ICCV 2017

ICCV 2017, Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks

- **Text + Keypoint --> Image**

GAN - NIPS 2016 ***PixelCNN - ICLRw 2017***

A man in a orange jacket with sunglasses and a hat ski down a hill.



NIPS 2016, Learning What and Where to Draw

ICLRw, 2017GENERATING INTERPRETABLE IMAGES WITH CONTROLLABLE STRUCTURE

Related Work

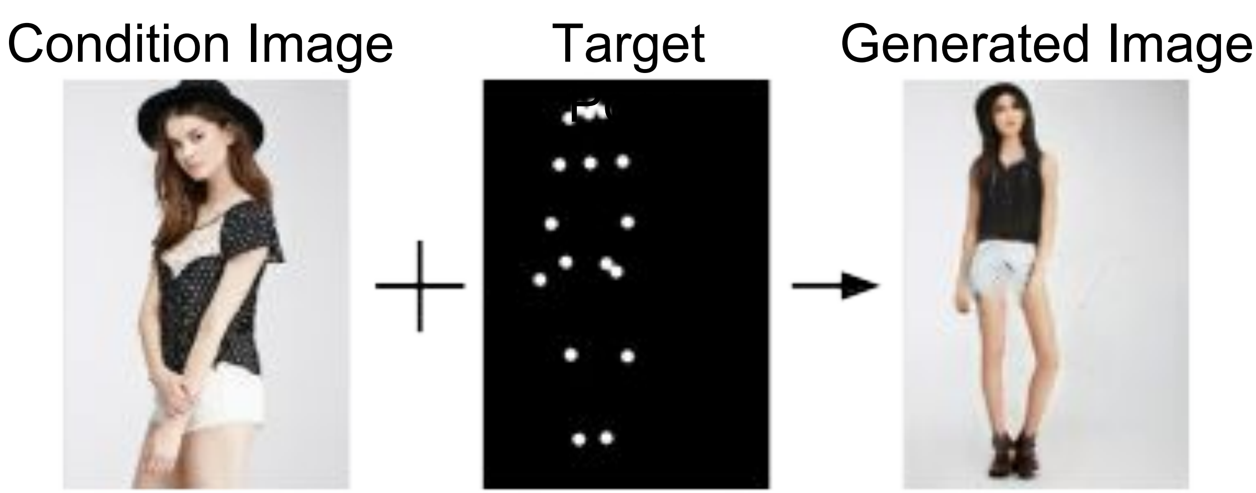
- **Image + Viewpoint --> Image**

VariGAN - Arxiv 2017

Arxiv 2017, Multi-View Image Generation from a Single-View

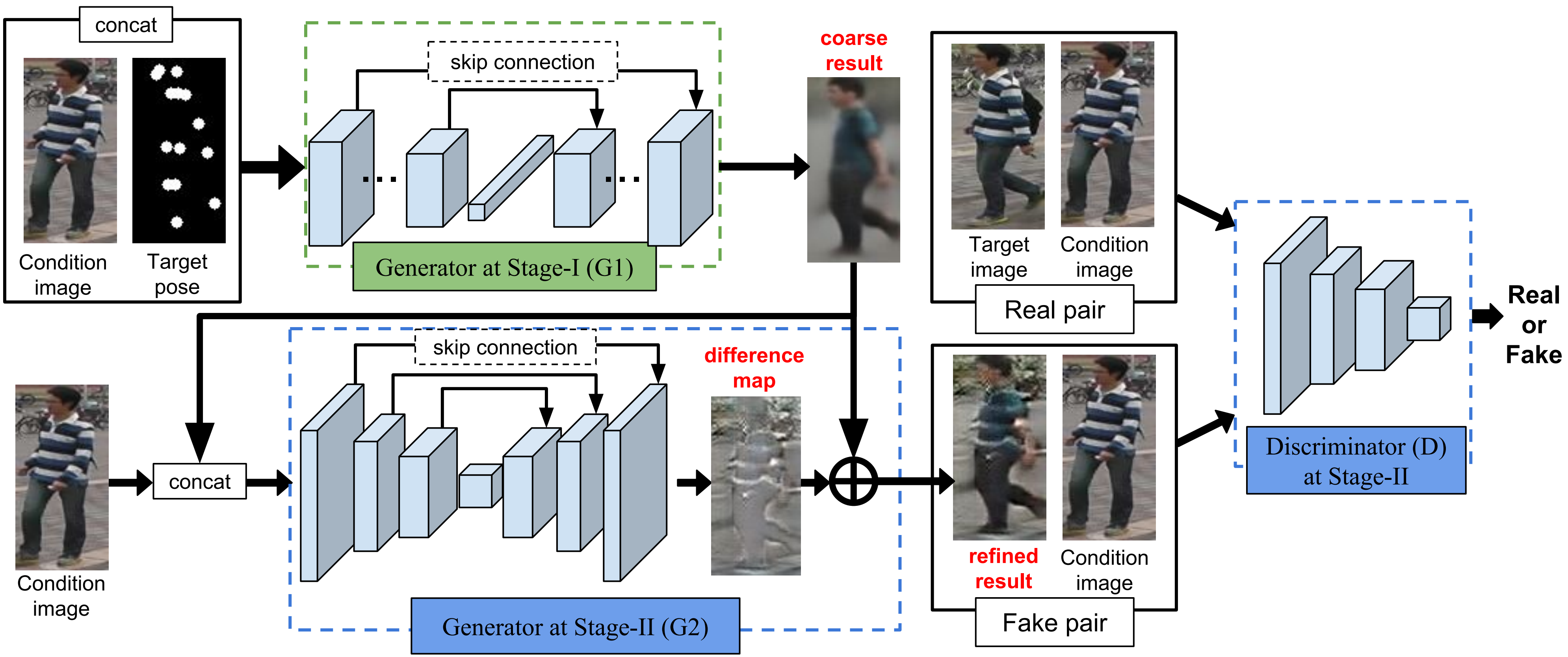
- **Our work**

More concrete appearance and structure information



Method

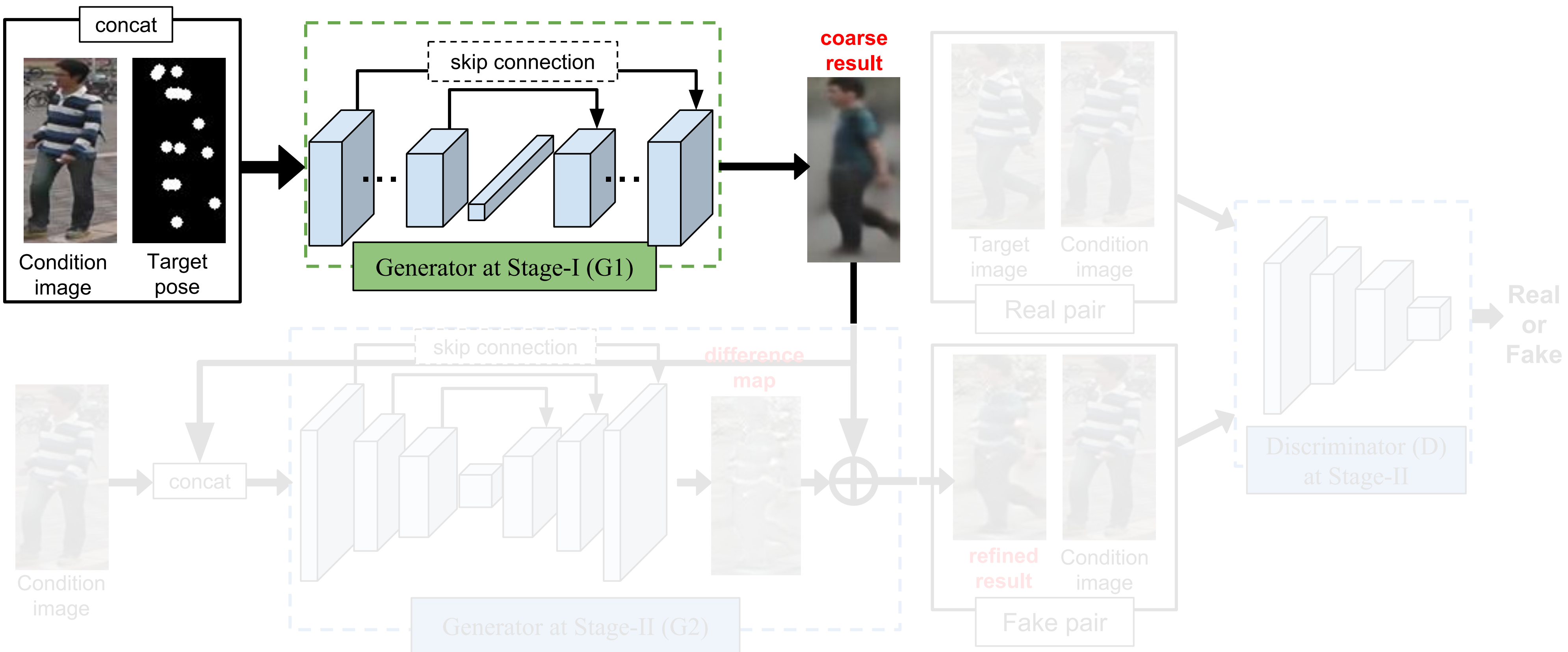
- **Our Two-stage Framework**



The overall framework of our Pose Guided Person Generation Network (PG2). It contains two stages focusing on pose and appearance, respectively.

Method

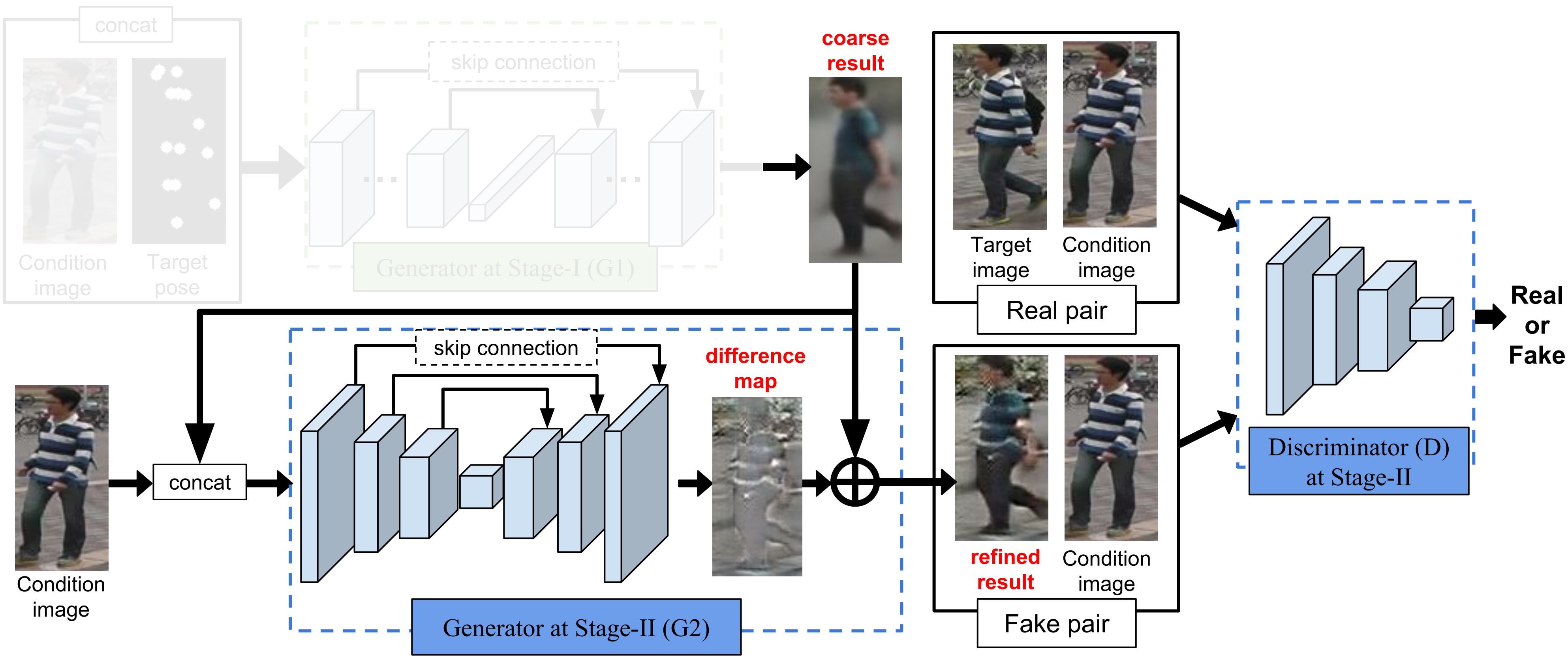
- **Our Two-stage Framework**



Stage-I focuses on pose integration and generates an initial result that captures the global structure of the human.

Method

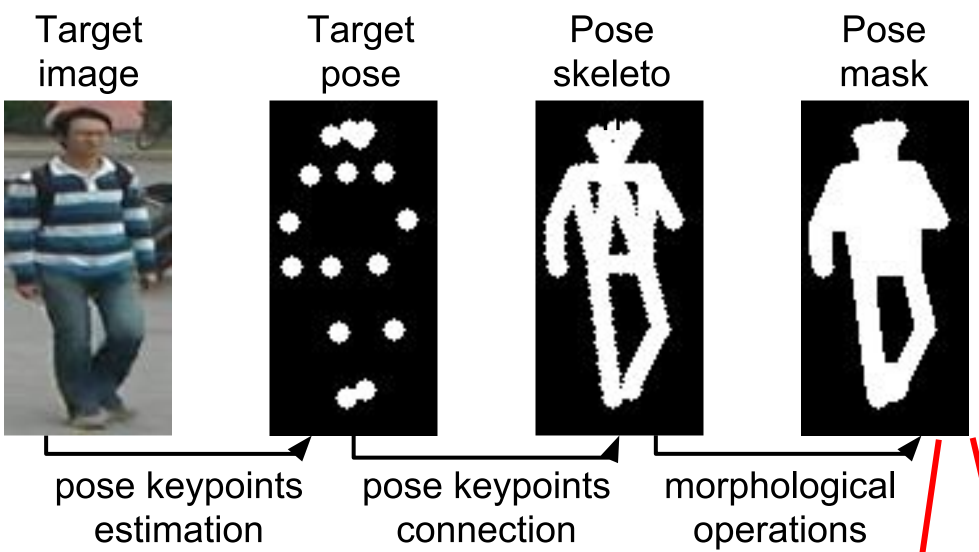
- **Our Two-stage Framework**



Stage-II focuses on refining the initial result via adversarial training and generates sharper images.

Method

- **Optimization losses**



Process of computing the pose mask

Stage-I:

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

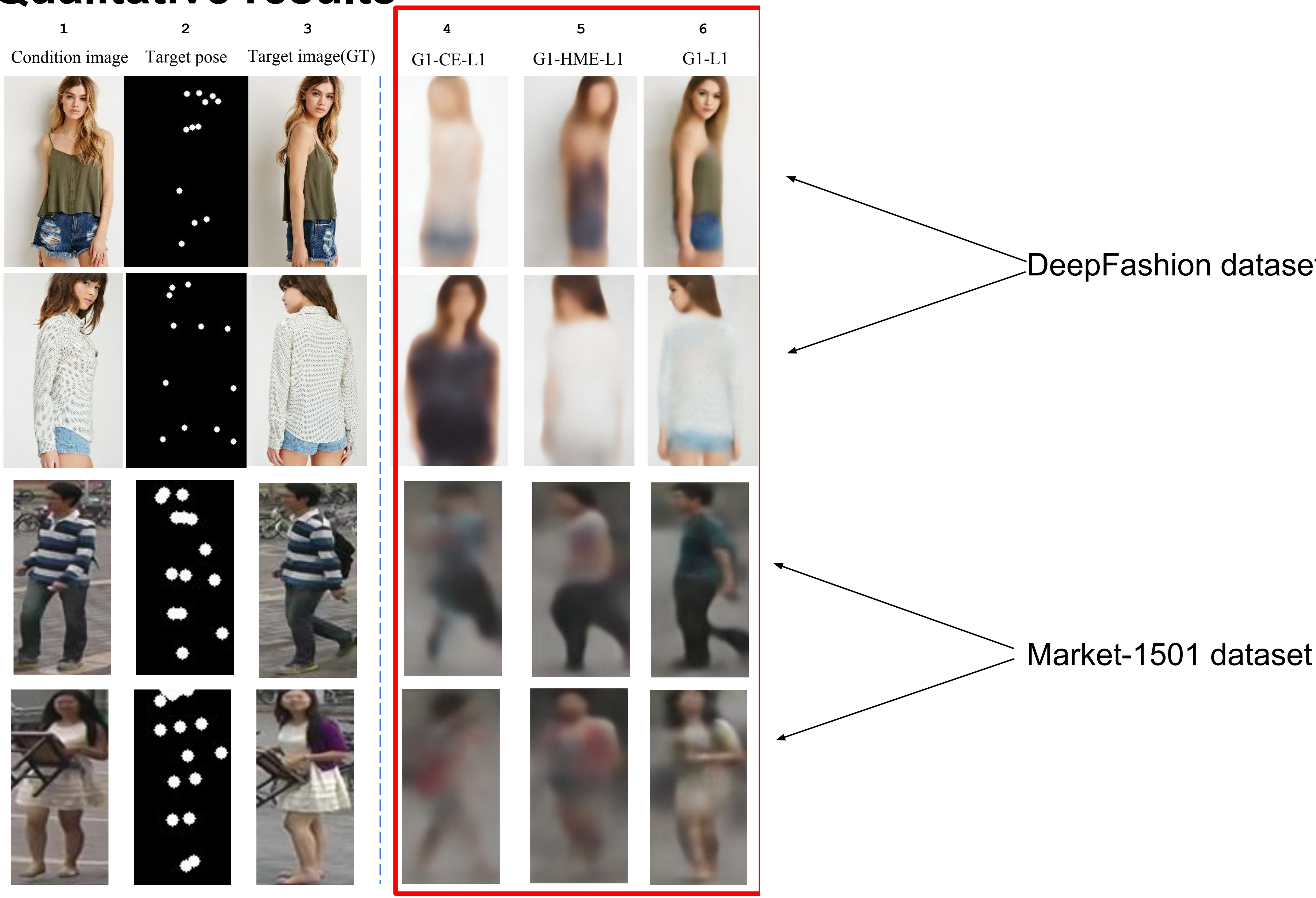
Stage-II:

$$\begin{aligned}
\mathcal{L}_{adv}^{\mathsf{D}} &= \mathcal{L}_{bce}(\mathsf{D}(I_A, I_B), 1) + \mathcal{L}_{bce}(\mathsf{D}(I_A, \mathsf{G2}(I_A, \hat{I}_{B1})), 0), \\
\mathcal{L}_{adv}^{\mathsf{G}} &= \mathcal{L}_{bce}(\mathsf{D}(I_A, \mathsf{G2}(I_A, \hat{I}_{B1})), 1), \\
\mathcal{L}_{\mathsf{G2}} &= \mathcal{L}_{adv}^{\mathsf{G}} + \lambda\|(\mathsf{G2}(I_A, \hat{I}_{B1}) - I_B) \odot (1 + M_B)\|_1,
\end{aligned}$$

Experiments

- **Qualitative results**

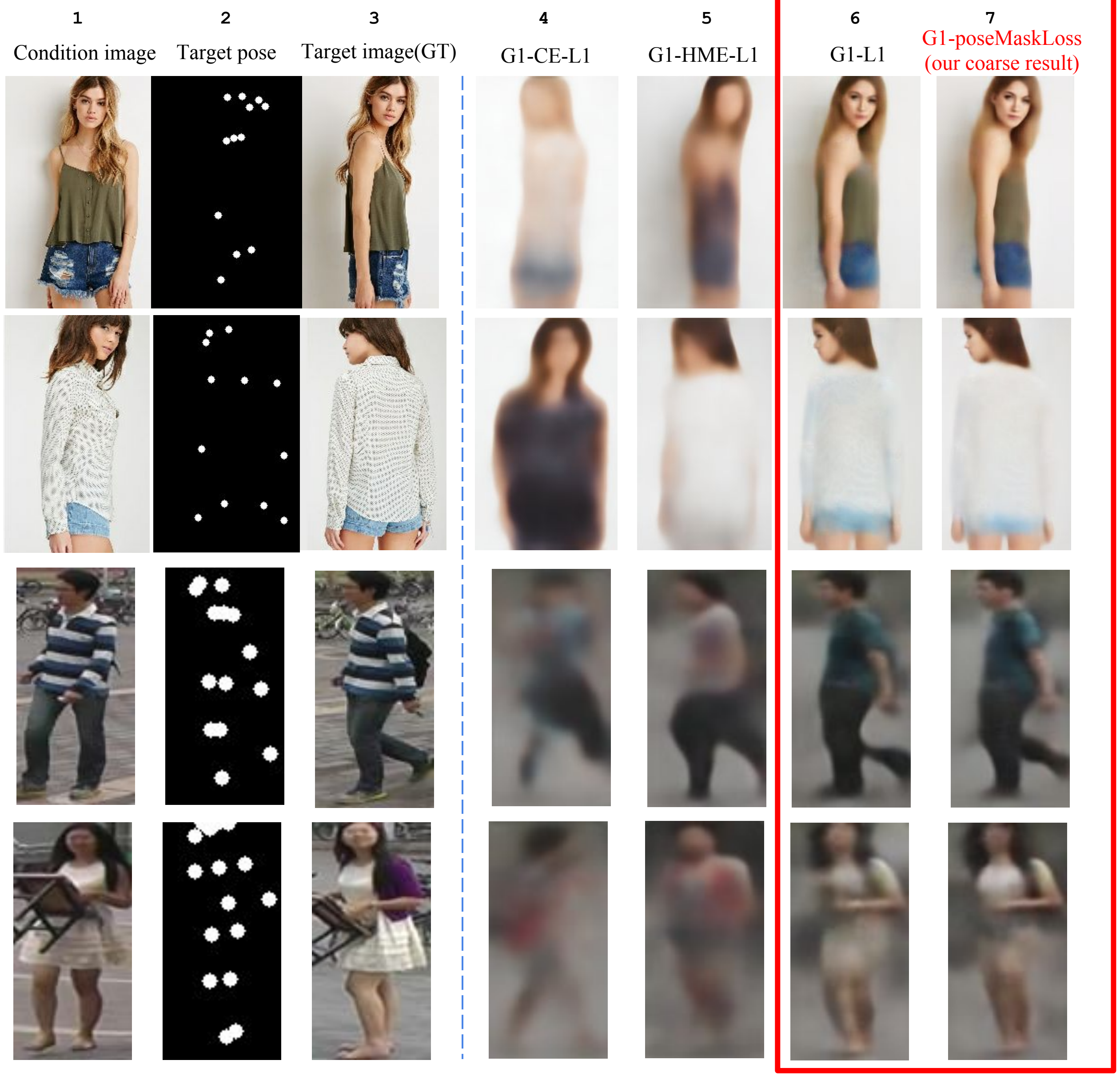
The proposed embedding method generates more accurate and realistic results than CE (Coordinate Embedding) and HME (Heat Map Embedding).



Experiments

- **Qualitative results**

The proposed Posemask loss results in sharper results.



Experiments

- **Qualitative results**

The proposed two-stage framework generate better results than one-stage.

Sharper arm and face

More texture

More texture and background details

Sharper legs and more object details

Experiments

- **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

- The proposed pose embedding (G1-L1) consistently outperforms G1-CE-L1 across all measures and both datasets. G1-HME-L1 obtains similar quantitative numbers probably due to the similarity of the two embeddings.
- Changing the loss from L1 to the proposed poseMaskLoss (G1-poseMaskLoss) consistently improves further across all measures and for both datasets.
- Adding the discriminator during training either after the first stage (G1+D) or in our full model (G1+G2+D) leads to comparable numbers, even though we have observed clear differences in the qualitative results as discussed above. This is explained by the fact that blurry images often get good SSIM despite being less convincing and photo-realistic.

Note: mask-SSIM and mask-IS to reduce the influence of background on Market-1501 dataset.

Experiments

- **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

- The proposed pose embedding (G1-L1) consistently outperforms G1-CE-L1 across all measures and both datasets. G1-HME-L1 obtains similar quantitative numbers probably due to the similarity of the two embeddings.
- Changing the loss from L1 to the proposed poseMaskLoss (G1-poseMaskLoss) consistently improves further across all measures and for both datasets.
- Adding the discriminator during training either after the first stage (G1+D) or in our full model (G1+G2+D) leads to comparable numbers, even though we have observed clear differences in the qualitative results as discussed above. This is explained by the fact that blurry images often get good SSIM despite being less convincing and photo-realistic.

Note: mask-SSIM and mask-IS to reduce the influence of background on Market-1501 dataset.

Experiments

- **Quantitative results**

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

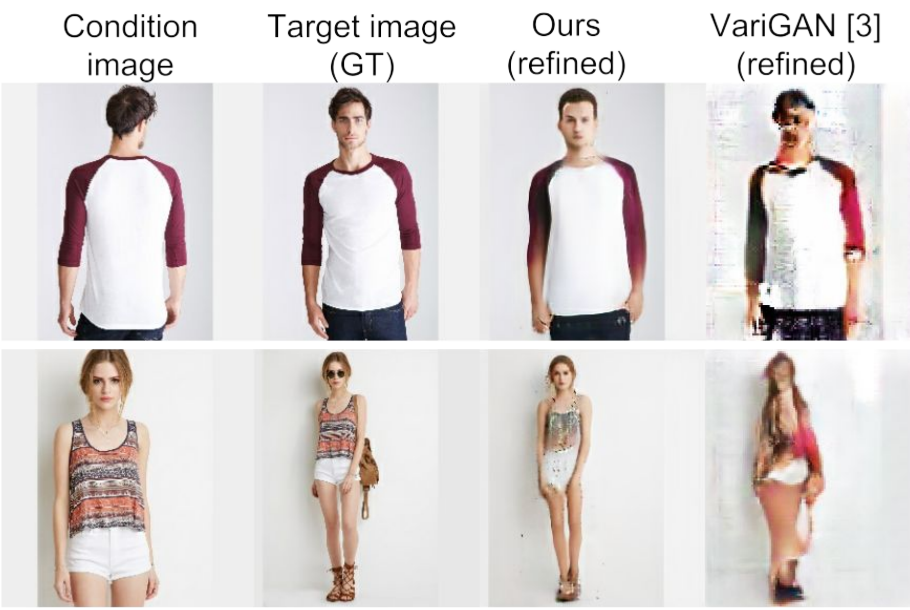
	DeepFashion		Market-1501			
Model	SSIM	IS	SSIM	IS	mask-SSIM	mask-IS
G1-CE-L1	0.694	2.395	0.219	2.568	0.771	2.455
G1-HME-L1	0.735	2.427	0.294	3.171	0.802	2.508
G1-L1	0.735	2.427	0.304	3.006	0.809	2.455
G1-poseMaskLoss	0.779	2.668	0.340	3.326	0.817	2.682
G1+D	0.761	3.091	0.283	3.490	0.803	3.310
G1+G2+D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

- The proposed pose embedding (G1-L1) consistently outperforms G1-CE-L1 across all measures and both datasets. G1-HME-L1 obtains similar quantitative numbers probably due to the similarity of the two embeddings.
- Changing the loss from L1 to the proposed poseMaskLoss (G1-poseMaskLoss) consistently improves further across all measures and for both datasets.
- Adding the discriminator during training either after the first stage (G1+D) or in our full model (G1+G2+D) leads to comparable numbers, even though we have observed clear differences in the qualitative results as discussed above. This is explained by the fact that blurry images often get good SSIM despite being less convincing and photo-realistic.

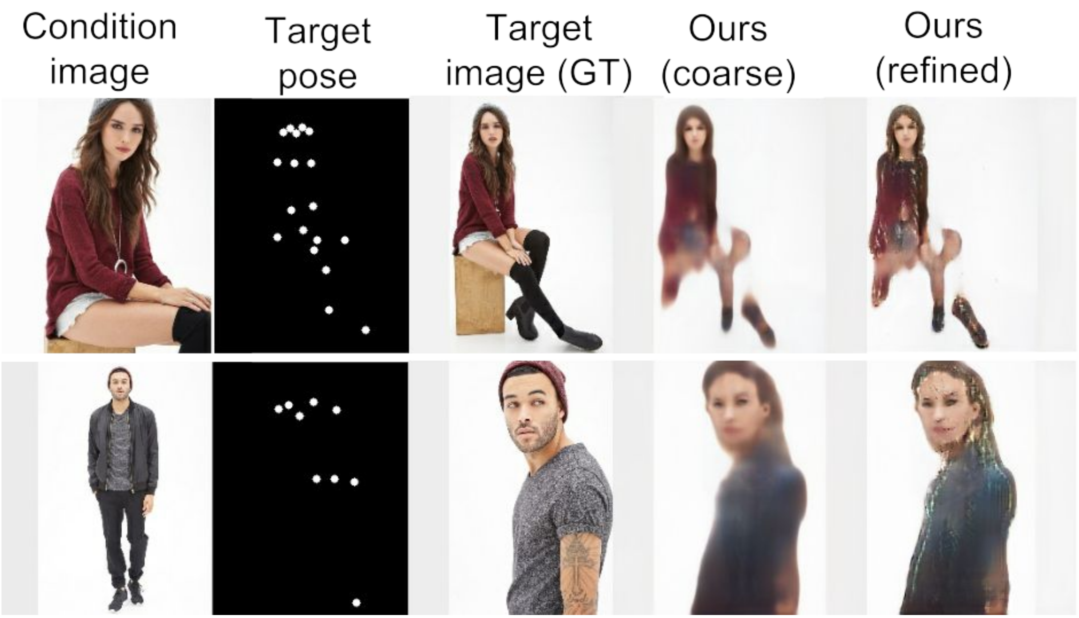
Note: mask-SSIM and mask-IS to reduce the influence of background on Market-1501 dataset.

Experiments

• Further analysis



Comparison to VariGAN

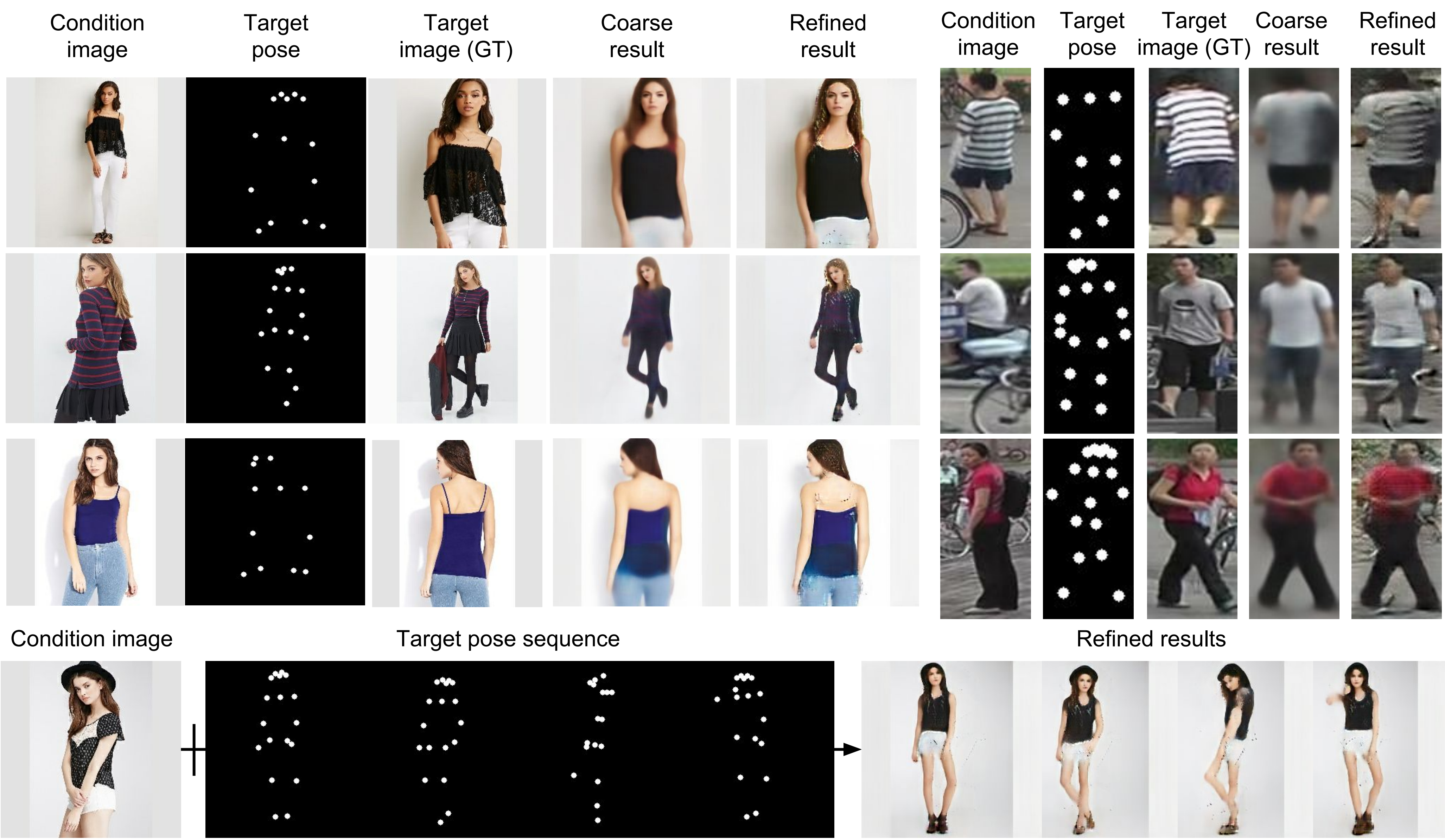


Failure cases on DeepFashion

- More realistic results, especially the faces.
- Generate whole body from half body.

- Rare data for some specific poses.
- Rare data for male.

Experiments

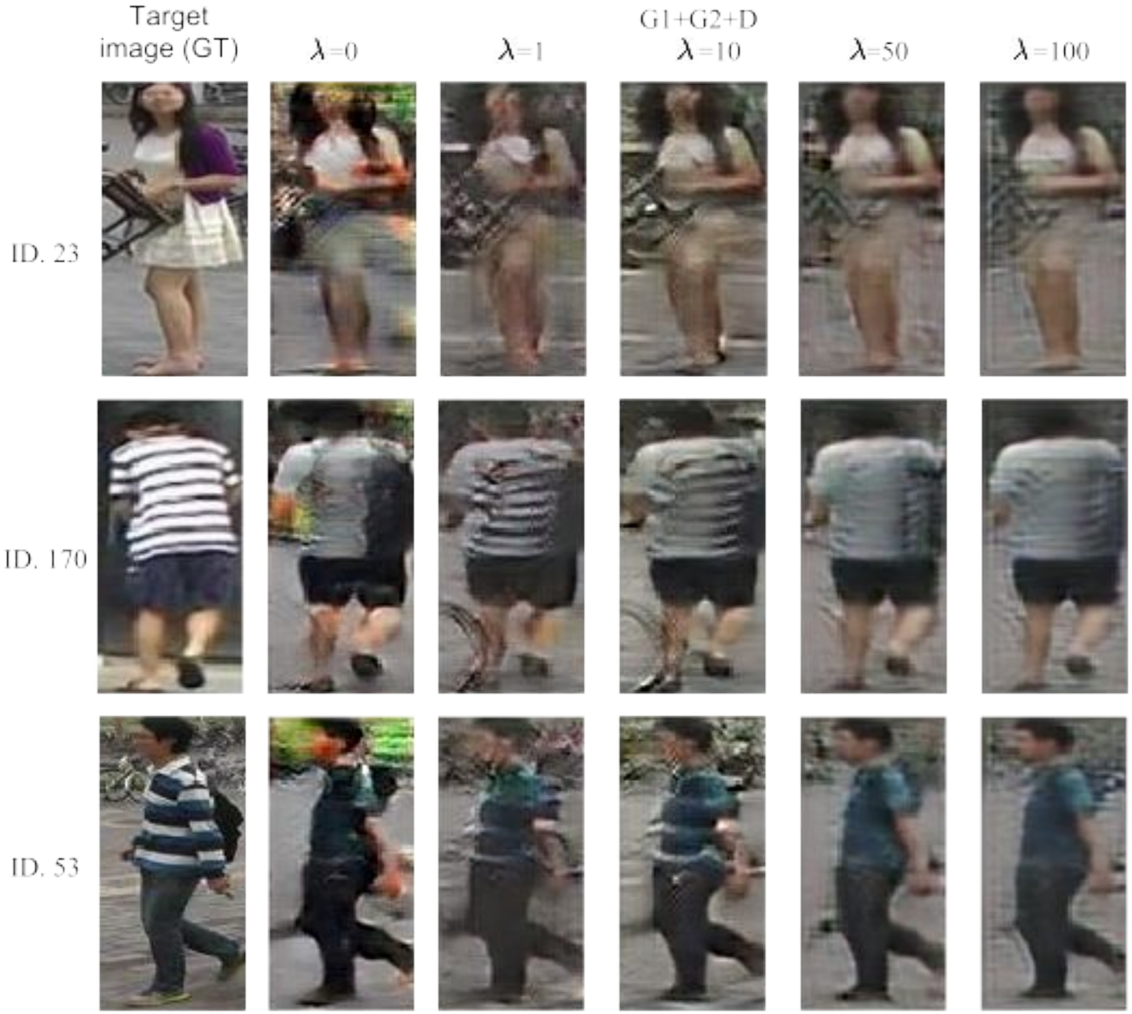


- Generate whole body from up body
- Generate front view from side view

Experiments

- **Influence of λ**

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



- smaller λ leads to more details and sharper images (except λ = 0)
- larger λ leads to less details and blurrier images

Conclusions

Contributions

- We propose a novel task of conditioning image generation on a reference image and an intended pose.
- We propose a two stages framework focusing on global body structure and local appearance details.
- Our method can be useful for several tasks (see Further Reading).

Further Reading

- Person re-identification

1) A Pose-Sensitive Embedding for Person Re-Identification with Expanded Cross Neighborhood Re-Ranking
2) Pose-Normalized Image Generation for Person Re-identification
3) Disentangled Person Image Generation

- Video Prediction

1) Deep Video Generation, Prediction and Completion of Human Action Sequences

- Face generation

1) Natural and Effective Obfuscation by Head Inpainting
2) Every Smile is Unique: Landmark-Guided Diverse Smile Generation

Questions ?