Pose Guided Person Image Generation

Liqian Ma¹, Xu Jia²*, Qianru Sun³*, Bernt Schiele³, Tinne Tuytelaars², Luc Van Gool¹,⁴
¹KU-Leuven/PSI, TRACE (Toyota Res in Europe)  ²KU-Leuven/PSI, IMEC
³Max Planck Institute for Informatics, Saarland Informatics Campus  ⁴ETH 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.

---

Condition Image  +  Target Pose  →  Generated Image
Related Work

- **Image --> Image**
  - Labels to Street Scene
  - Day to Night
  - 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**
  - Key-points
  - GAN - NIPS 2016
  - A man in a orange jacket with sunglasses and a hat ski down a hill.
  - NIPS 2016, Learning What and Where to Draw
  - PixelCNN - ICLRw 2017
  - ICLRw, 2017 GENERATING 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
The overall framework of our Pose Guided Person Generation Network (PG2). It contains two stages focusing on pose and appearance, respectively.
Stage-I focuses on pose integration and generates an initial result that captures the global structure of the human.
Stage-II focuses on refining the initial result via adversarial training and generates sharper images.
Method

- Optimization losses

![Diagram showing the process of computing the pose mask]

\[
\mathcal{L}_{G1} = \| (G_1(I_A, P_B) - I_B) \odot (1 + M_B) \|_1
\]

Stage-I:

\[
\mathcal{L}_{adv}^D = \mathcal{L}_{bce}(D(I_A, I_B), 1) + \mathcal{L}_{bce}(D(I_A, G_2(I_A, \hat{I}_B)), 0),
\]

\[
\mathcal{L}_{adv}^G = \mathcal{L}_{bce}(D(I_A, G_2(I_A, \hat{I}_B)), 1),
\]

\[
\mathcal{L}_{G2} = \mathcal{L}_{adv}^G + \lambda \| (G_2(I_A, \hat{I}_B) - I_B) \odot (1 + M_B) \|_1,
\]
Experiments

• **Qualitative results**

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

   - **DeepFashion dataset**

   - **Market-1501 dataset**
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.

1. Condition image
2. Target pose
3. Target image (GT)
4. G1-CE-L1
5. G1-HME-L1
6. G1-L1
7. G1-poseMaskLoss (our coarse result)
8. G1+D (our refined result)
9. G1+G2+D

- Sharper arm and face
- More texture
- More texture and background details
- Sharper legs and more object details
• Quantitative results

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

| Model                  | DeepFashion | Market-1501 |
|------------------------|-------------|-------------|
|                         | 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.

| Model               | DeepFashion | Market-1501 |
|---------------------|-------------|-------------|
|                     | 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.
• **Quantitative results**

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

| Model               | DeepFashion | Market-1501 |
|---------------------|-------------|-------------|
|                     | 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**
- More realistic results, especially the faces.
- Generate whole body from half body.

**Failure cases on DeepFashion**
- 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 $\lambda$

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

- smaller $\lambda$ leads to more details and sharper images (except $\lambda = 0$)
- larger $\lambda$ 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 ?