Neural Techniques for Pose Guided Image Generation

Ayush Gupta (ayushg@stanford.edu), Shreyash Pandey (shreyash@stanford.edu), Vivekkumar Patel (vivek14@stanford.edu)

2018

Introduction

- To sythesize person images based on a condition image and desired pose.
- Extend the PG^2 model [3] for better performance and speedier training.
- Incorporate WGANs for robust and more stable training.
- Use a triplet classification based discriminator architecture for model simplification and speed.

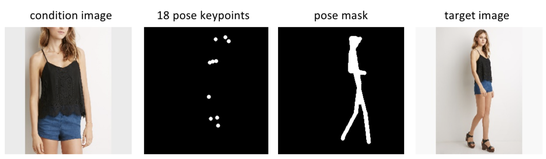
Data Preparation

Dataset

- The DeepFashion (In-shop Clothes Retrieval Benchmark) dataset is used.
- It consists of 52,712 number of in-shop clothes images and around 200,000 cross-pose/scale pairs.
- Each image has a resolution of 256 x 256.

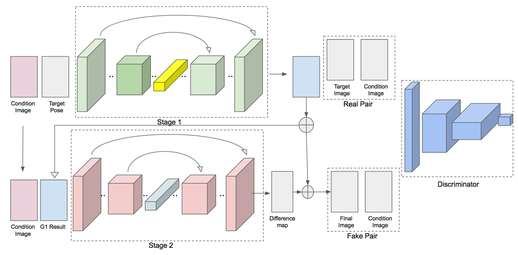
Processing

- Filtered for non-single and complete images. Split into 32,031 training and 7,996 test images.
- Training set augmented to 57,520 images after left-right flip and filtering non-different images.
- Gives **127,022 training ex.; 18,568 test ex.**
- A state-of-the-art pose estimator [2] gives 18 pose keypoints of images. Stored as pickle files.
- The keypoints were also used to generate a human body mask by morphological transformations for Stage-1 training



- Thus, each example stored in HDF5 contained a condition image, target image, 18 channel target pose info, and a single channel pose mask.
- Each image pair was normalized by a mean of 127.5, and std. of 127.5.
- Poses normalized to be between -1 and +1.

Original Architecture



- Stage-1 has a U-net like architecture. Skip connections help in transferring appearance information to the output.
- Stage-2 is a conditional DCGAN that generates a difference map.
- Stage-2 Discriminator differentiates between pairs of images. Real pair has condition and target image. Fake pair has condition and generated image.

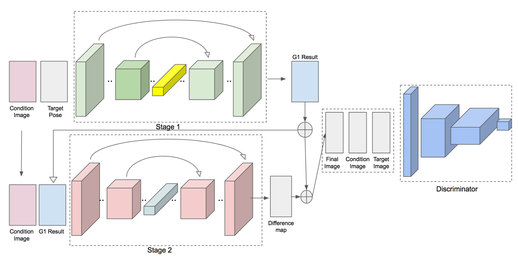
Pose Mask Loss

- We calculate the loss for stage-1 as follows:

$$L = ||(\hat{I}_B - I_B) * (1 + M_B)||_1$$

- Masking by pose forces the model to ignore the background variations.

Our Modifications



- Discriminator is now similar to a classifier. Target image has label 1, generated image and condition image have label 0.
- Using WGAN training to train the conditional DCGAN in stage-2.
- With WGAN, Discriminator is trained for N (= 10) iterations for every iteration on Generator.

Implementation Details

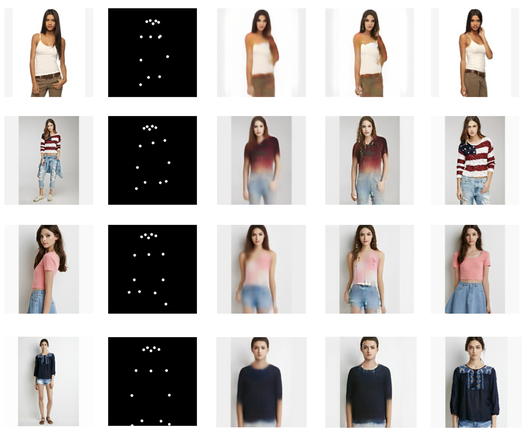
- With a batch-size of 4, Stage-1 is trained for 40k steps followed by 25k steps for Stage-2.
- Adam Optimizer is used for Stage-1 and Stage-2, except for WGAN in which we use RMSProp for Stage-2.
- Learning rates for each of these stages are set to 5e-5.

SSIM and IS scores

No.	Model	IS	SSIM
1	G_1	2.58	0.72
2	$G_1 + G_2 + D$	2.993	0.71
3	Triplet	3.01	0.727
4	**WGAN-triplet**	**3.02**	**0.73**
5	**Ma et. al.**	**3.091**	**0.76**
6	Target	3.25	1

Table 1: Ma et al. was the original paper. Target denotes the target test set.

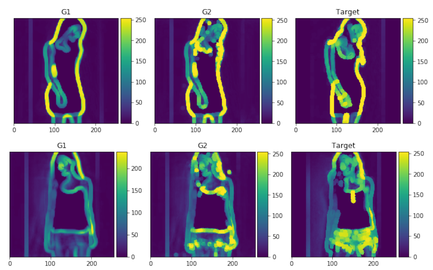
Generation Results



- Shown above are results of WGAN-triplet model in the following order: from left to right, condition image, target pose, Stage-1 result, Stage-2 result and target image.
- Stage-1 combines the target pose and appearance.
- Stage-2 generates a sharper image after using the coarse result.

Failures

Sharpness visualized via Gradients



- Stage-2 helps in generating sharper images.

Conclusion and Future Work

- Two-stage architecture is able to transfer pose while retaining appearance.
- Incorporating triplet classification and WGAN leads to faster convergence and more robust training.
- Failure modes include target poses that involve closeups of people, and attires such as jackets that look different from different angles. More data could potentially solve these problems.
- One extension could be to extend this model to a semi-supervised setting where pose is supervised and appearance is unsupervised. This will allow arbitrary manipulation of poses [1].

References

[1] Rodrigo de Bem et al. "DGPose: Disentangled Semi-supervised Deep Generative Models for Human Body Analysis". In: *arXiv preprint arXiv:1804.06364* (2018).

[2] Zhe Cao et al. "Realtime multi-person 2d pose estimation using part affinity fields". In: *CVPR*. Vol. 1. 2. 2017, p. 7.

[3] Liqian Ma et al. "Pose guided person image generation". In: *Advances in Neural Information Processing Systems*. 2017, pp. 405–415.