Multiview Human Synthesis From a Single View

Si Wen, Tiancong Zhou, Honghao Qiu
{wensi, longztc, honq}@stanford.edu

Abstract/Motivation

We use generative deep learning models to synthesize multiview images given a single view. One potential application is to generate multiview images for e-Commerce products. The generation process is done in two stages: in the first stage, we train a variational auto-encoder (VAE) to synthesize a new view of the input image; in the second stage, we use a generative adversarial network (GAN) to generate details on the output of the first stage. We evaluate our results using both qualitative and quantitative methods.

Dataset

Data: We used a combination of MVC datasets (160,000 real images) and 100,000 synthetic images (360-degrees views). Image data is labeled with an angle (e.g. 136 degree). Examples of images in the synthetic dataset:

Future Work

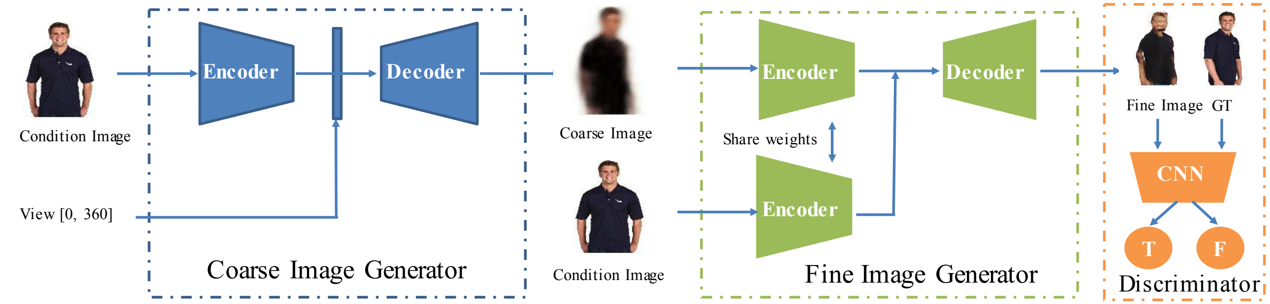
1. Improve the GAN: we only used a simple conditional GAN with very little hyperparameter tuning. There is a number of recently published papers that shows various techniques to improve the quality and stability of GANs.
2. Improve face synthesis: our model currently handles face poorly since it contains a lot of noticeable details. It is important to be able to recreate these details in a convincing way for good result.
3. Include background: our current dataset only contains images without any background. It is more difficult to synthesize views when the object of interest is not presented in isolation and we would like to tackle that problem in the future.

References

[1] Goodfellow et al. Generative Adversarial Nets, *NIPS 2014*.
[2] M. Mirza and S. Osindero. Conditional Generative Adversarial Nets. *arXiv:1411.1784, 2014*.
[3] D. P. Kingma and M. Welling. Auto-encoding Variational bayes. *ICLR, 2014*.
[4] Salimans et al. Improved Techniques for Training GANs. *arXiv:1606.03498, 2016*.
[5] Zhao et al. Multi-View Image Generation from a Single-View. *arXiv 1704.04886, 2017*.

Our Approach

➢ **Network Architecture:**



The VAE encodes the input image, concatenates it with the target angle (scalar value in [0, 360]), and feeds them through the decoder to produce an image in the target view.

We feed the input image and the output from VAE through a Siamese network to produce the final fine image in the target view, and is trained adversarially.

➢ **Objective (Loss):**

VAE Loss:

$$L(\theta;x) = -KL(q_\theta(z|x)||p(z)) + E_{q_\theta(z|x)}[log(p_\theta(x|z))]$$

Adversarial Loss:

$$E_{I_v \sim p_{data(I_v)}}[logD(I_v)] + E_{z\sim p(z)}[log(1 - D(I_v, G(z, I_v)))]$$

Experimental Results

We use several different quantitative methods to evaluate the quality of our model, including the Structural Similarity Index (SSIM) and the Inception Score (IS).

Methods	SSIM	IS	RMSE
VAE	0.59 ± 0.10	1.67 ± 0.30	7.49 ± 0.40
GAN	0.63 ± 0.09	2.62 ± 0.38	6.70 ± 0.58
Ours	0.66 ± 0.10	2.35 ± 0.45	6.62 ± 0.69

$$SSIM(I_x, I_y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \quad , \quad IS(I_x, I_y) = exp(E_{I_x} D_{KL(p(y|I_x)||p(y))})$$

where I_x, I_y are generated image and GT image respectively, y is the label (the input angle in our model).

Discussions

1. Our model works well on synthetic images, as there are less noise in the data (e.g. no background, unified pose, similar texture, etc.).
2. On MVC dataset, we achieved reasonably good result compared to other methods (see table above). Since VAE is capable of finding global appearance with less details, while GAN is good at filling rich details of the synthesized image but less capable of capturing global appearance and rough outlines for human/clothings, therefore, GAN and VAE complements each other well for image synthesis in our architecture. The coarse image generated by VAE captures global appearance, while GAN fills in details to make the fine image.
3. Our model does not perform well on face synthesis (compared to the body), since face contains more details and is harder to synthesize.