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

Our Approach

Network Architecture:
Encoder | Decoder
Condition Image
Coarse Image Generator
View [0, 360]

Our Approach

Objective (Loss):
VAE Loss:
\[
L(\theta; x) = -KL(q(y|x)||p(z)) + E_{q(y|x)}[\log(p_{\theta}(x|y))]
\]

Adversarial Loss:
\[
E_L = p_{data}(I_c)[\log D(I_c)] + E_{L_{GAN}}(I_c) = \exp(E_{D}(E_{L_{GAN}}(I_c)))
\]

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).

<table>
<thead>
<tr>
<th>Methods</th>
<th>SSIM</th>
<th>IS</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>VAE</td>
<td>0.59 ± 0.10</td>
<td>1.67 ± 0.30</td>
<td>7.49 ± 0.40</td>
</tr>
<tr>
<td>GAN</td>
<td>0.63 ± 0.09</td>
<td>2.62 ± 0.38</td>
<td>6.70 ± 0.58</td>
</tr>
<tr>
<td>Ours</td>
<td>0.66 ± 0.10</td>
<td>2.35 ± 0.45</td>
<td>6.62 ± 0.69</td>
</tr>
</tbody>
</table>

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.