ST-GAN: Unsupervised Facial Image Semantic Transformation Using Generative Adversarial Networks

Appendix A. The objective functions of WGAN-GP

For D network:

$$\begin{aligned} \min_D L_{GAN} &= \mathbb{E}_{z\sim P_z(z),c\sim P_c(c)}[D(G(z,c))] - \mathbb{E}_{x\sim P_{data}(x)}[D(x)] \\ &+ \lambda\mathbb{E}(t), \end{aligned} \tag{1}$$

where $t = (\|\nabla_{\hat{x}} D(\hat{x})\|_2 - 1)^2$. Here, $\hat{x} = \epsilon x + (1-\epsilon)\tilde{x}$, where $\epsilon \sim U[0,1]$, $x \sim P_{data}(x)$, $\tilde{x} \sim P_G(x)$.

For G network:

$$\min_G L_{GAN} = \mathbb{E}_{x\sim P_{data}(x)}[D(x)] - \mathbb{E}_{z\sim P_z(z),c\sim P_c(c)}[D(G(z,c))] \tag{2}$$

Appendix B. Mutual information term

The mutual information term $I(c; G(z,c))$ requires the posterior $P(c|G(z,c))$, thus, it is hard to maximize directly. ST-GAN uses a technique called Variational Information Maximization Barber and Agakov (2003) by defining an auxiliary distribution $Q(c|x)$ to approximate $P(c|x)$ as InfoGAN Chen et al. (2016) does. The variational lower bound, $L_I(G,Q)$, of the local mutual information $I(c; G(z,c))$ is defined as:

$$\begin{aligned} L_I(G,Q) &= \mathbb{E}_{c\sim P(c),x\sim G(z,c)}[logQ(c|x)] + H(c) \\ &= \mathbb{E}_{x\sim G(z,c)}[\mathbb{E}_{c'\sim P(c|x)}[logQ(c'|x)]] + H(c) \\ &\leq I(c; G(z,c)) \end{aligned} \tag{3}$$

ST-GAN simply adds some fully connected layers to D and the output of the final layer is regarded as the parameters of conditional distribution $Q(c|x)$. Finally, we replace the term $I(c; G(z,c))$ with $L_I(G,Q)$ for the objective function of ST-GAN, thus, the practical objective functions for G and D of ST-GAN is:

$$\begin{aligned} \min_G L_{GAN} &= \mathbb{E}_{x\sim P_{data}(x)}[D(x)] - \mathbb{E}_{z\sim P_z(z),c\sim P_c(c)}[D(G(z,c))] \\ &- \lambda_2 L_I(G,Q) \end{aligned} \tag{4}$$

$$\begin{aligned} \min_D L_{GAN} &= \mathbb{E}_{z\sim P_z(z),c\sim P_c(c)}[D(G(z,c))] - \mathbb{E}_{x\sim P_{data}(x)}[D(x)] \\ &+ \lambda_1\mathbb{E}(t) - \lambda_2 L_I(G,Q) \end{aligned} \tag{5}$$

The practical objective functions for LST-GAN are the same as ST-GAN, except replacing $Q(c|x)$ of $L_I(G,Q)$ with $Q(c|\tilde{x}_{local})$, where $\tilde{x}_{local} = F(G(z,c))$.

Table 1: Inception-scores for VAE/GAN and ST-GAN, evaluated on 12800 images. the variances of these scores are very small value, thus they has strong credibility.

Method	Inception Score
VAE/GAN Larsen et al. (2015)	2.80±0.057
ST-GAN	**2.86±0.04**
CelebA dataset	3.06±0.08

Appendix C. Assessment of image quality

In this experiment, we compared generated samples quality of ST-GAN with VAE/GAN Larsen et al. (2015). We trained separately every method on the CelebA training dataset and used Inception score Salimans et al. (2016) to evaluate the sample quality in 12800 images. The comparison results are shown in Table 1 and the Inception score of the CelebA dataset is performed in the last row of Table 1. Comparisons show ST-GAN get better generated results than VAE/GAN.

References

David Barber and Felix Agakov. The im algorithm: a variational approach to information maximization. In *Proceedings of the 16th International Conference on Neural Information Processing Systems*, pages 201–208. MIT Press, 2003.

Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2172–2180, 2016.

Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle, and Ole Winther. Autoencoding beyond pixels using a learned similarity metric. *arXiv preprint arXiv:1512.09300*, 2015.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Advances in Neural Information Processing Systems*, pages 2234–2242, 2016.