Fashion Generation from Text Descriptions Using AttnGAN

Liang Yang
lyang6@stanford.edu

Motivation

Text description given by a designer or a customer:
Short sleeve cotton polo in black

?

Challenge: Difficulty in obtaining high quality results for conditional GAN.

Purpose: This project is going to investigate applying AttnGAN for conditional generation of fashion products.

Data

Data : Image-text pairs. 113211 pairs for training, and 14148 pairs for validation.

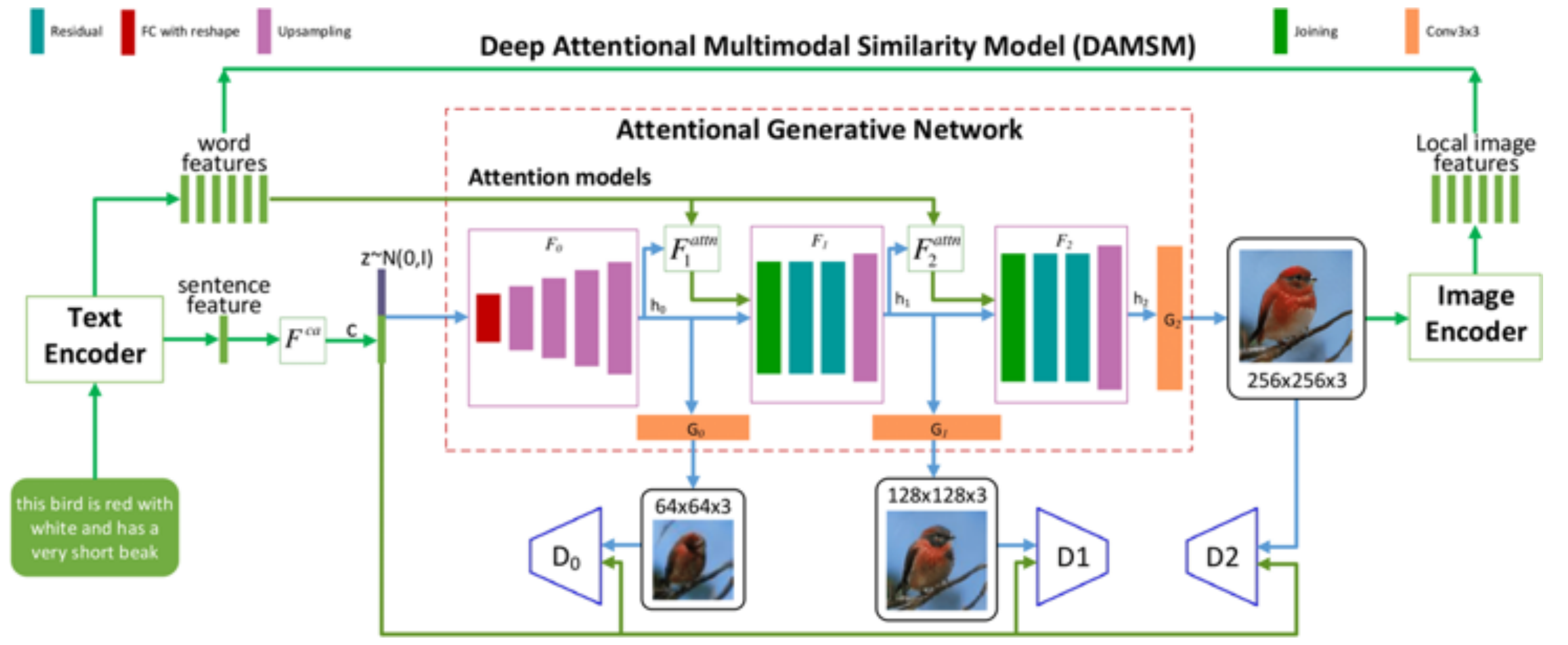
Text description:
Long sleeve flannel plaid shirt in tones of white and brown

Image

Methods

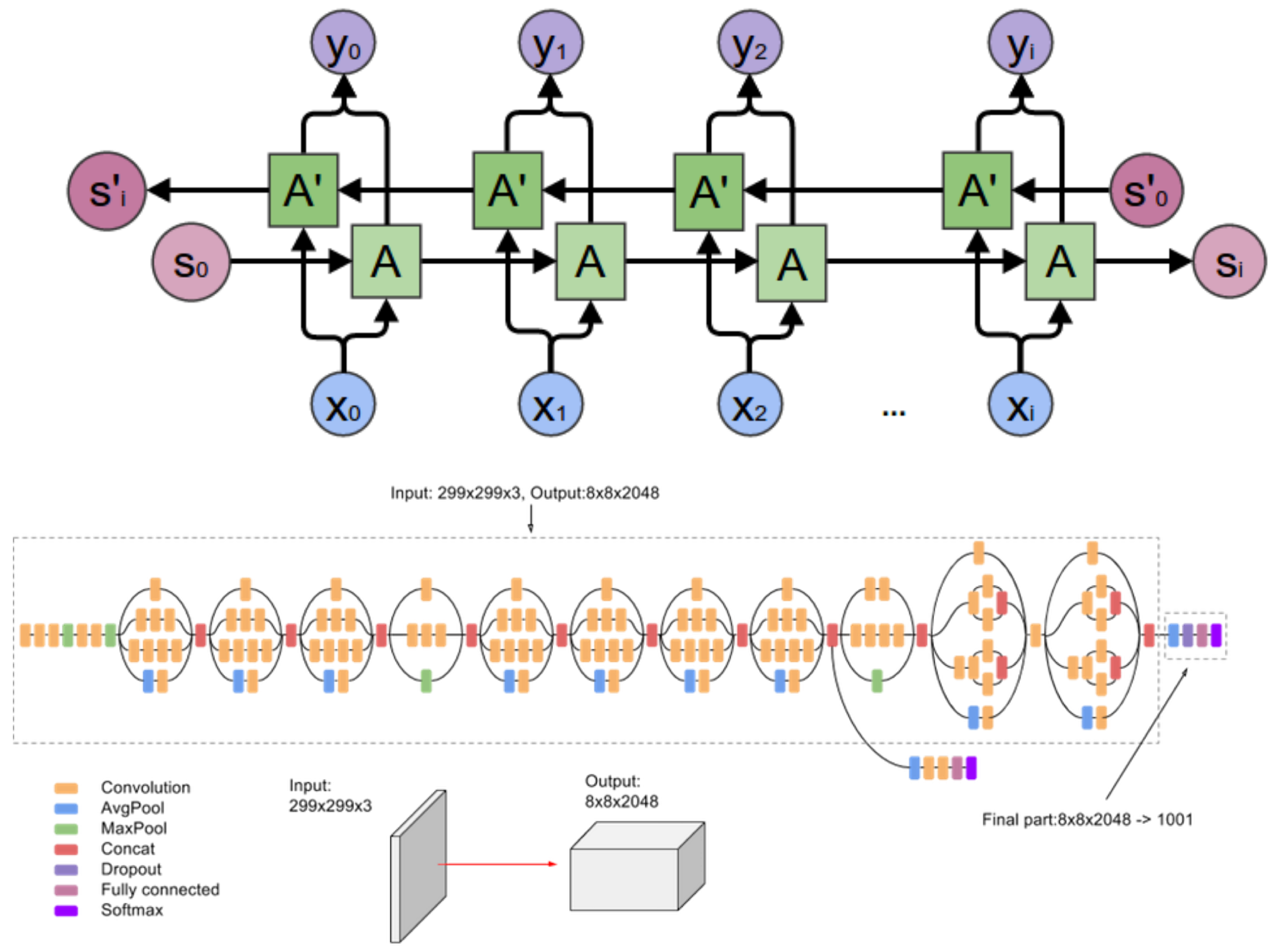
Generator loss: $\mathcal{L}_{G_i} = -\frac{1}{2}\mathbb{E}_{\hat{\boldsymbol{x}}_i \sim p_{G_i}}\left[\log\left(D_i(\hat{\boldsymbol{x}}_i)\right)\right] - \frac{1}{2}\mathbb{E}_{\hat{\boldsymbol{x}}_i \sim p_{G_i}}\left[\log\left(D_i(\hat{\boldsymbol{x}}_i, \bar{e})\right)\right]$

Discriminator loss: $\mathcal{L}_{D_i} = -\frac{1}{2}\mathbb{E}_{\boldsymbol{x}_i \sim p_{data_i}}\left[\log\left(D_i(\boldsymbol{x}_i)\right)\right] - \frac{1}{2}\mathbb{E}_{\hat{\boldsymbol{x}}_i \sim p_{G_i}}\left[1 - \log\left(D_i(\hat{\boldsymbol{x}}_i)\right)\right] - \frac{1}{2}\mathbb{E}_{\boldsymbol{x}_i \sim p_{data_i}}\left[\log\left(D_i(\boldsymbol{x}_i, \bar{e})\right)\right] - \frac{1}{2}\mathbb{E}_{\hat{\boldsymbol{x}}_i \sim p_{G_i}}\left[1 - \log\left(D_i(\hat{\boldsymbol{x}}_i, \bar{e})\right)\right]$



Attentional Generative Adversarial Network (AttnGAN) (Xu et al., 2018)

Methods



Bi-LSTM and CNN (Inception-v3) encoders used in DAMSM

Loss function: $\mathcal{L} = \mathcal{L}_G + \lambda\mathcal{L}_{DAMSM}$

Results

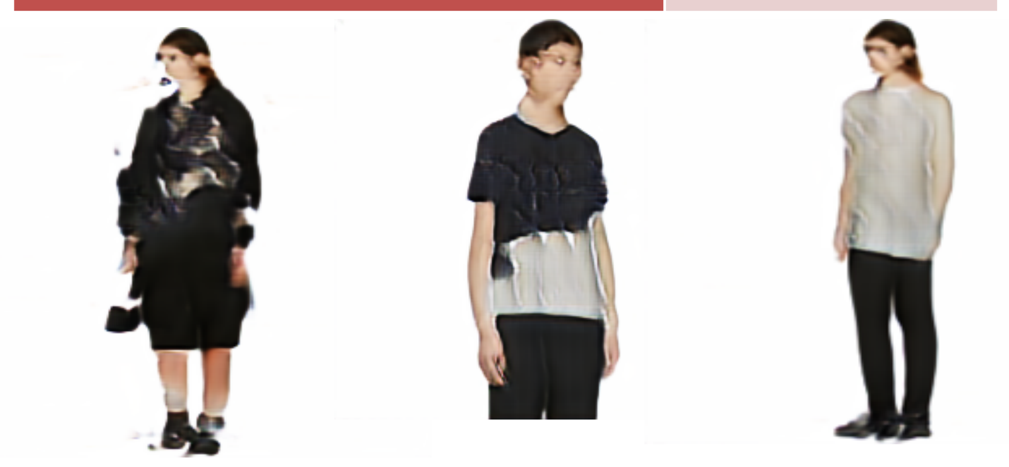
Performance measure:

Inception score: $IS(G) = \exp\left(\mathbb{E}_{\boldsymbol{x} \sim p_g} D_{KL}\left(p(y|\boldsymbol{x}) || p(y)\right)\right)$

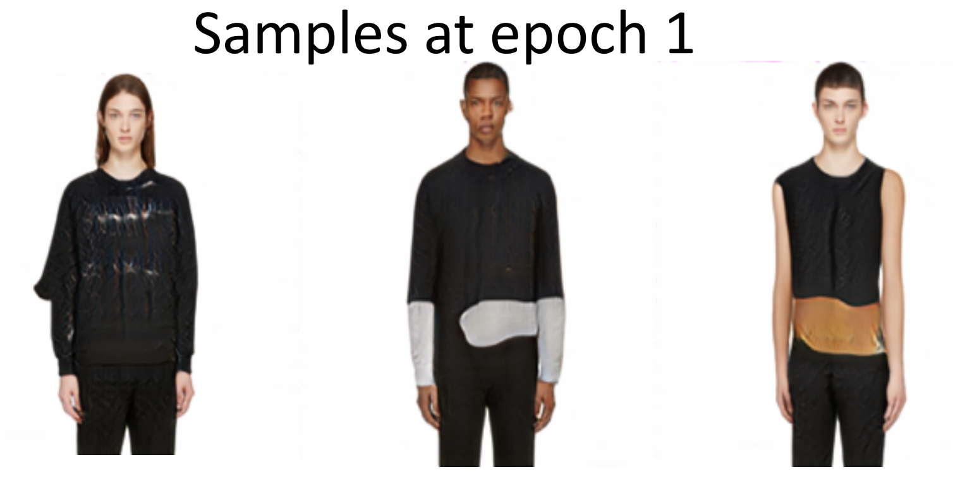
Table 1. Comparison of Hyperparameters

Hyperparameters	DAMSM loss after 10 epochs
$\boldsymbol{\gamma_1 = 5, \gamma_2 = 5, \gamma_3 = 10}$	$\boldsymbol{L^s = 6.87}$ $\boldsymbol{L^w = 6.79}$
$\gamma_1 = 5, \gamma_2 = 5, \gamma_3 = 50$	$L^s = 7.02$ $L^w = 6.88$
$\gamma_1 = 1, \gamma_2 = 1, \gamma_3 = 10$	$L^s = 7.01$ $L^w = 6.81$

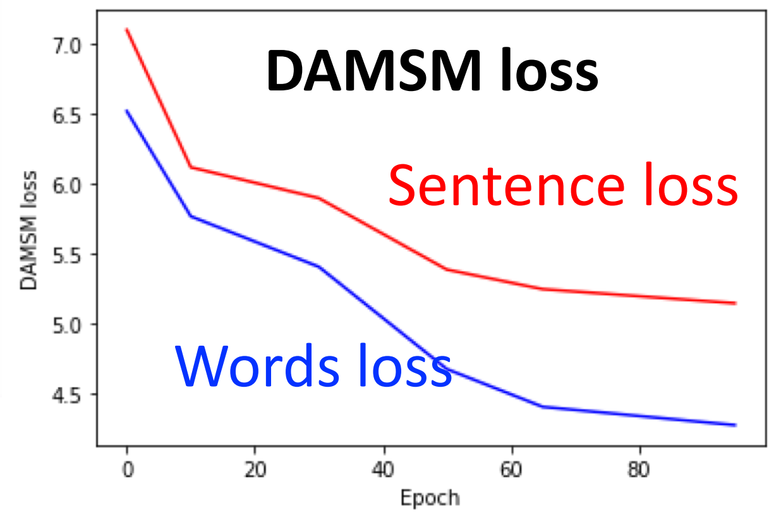
Inception Scores	$\lambda = 0$	$\lambda = 1$	$\boldsymbol{\lambda = 10}$	$\lambda = 50$
Mean	2.61	2.77	**3.03**	2.42
Stdv	0.13	0.20	**0.18**	0.20

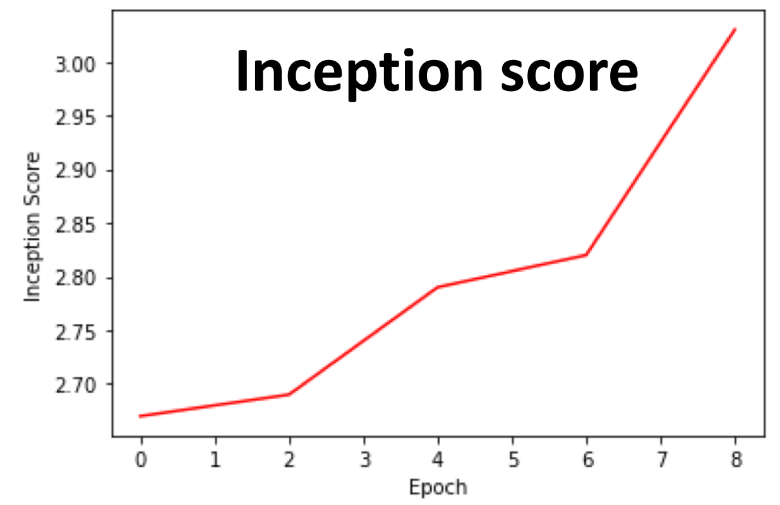


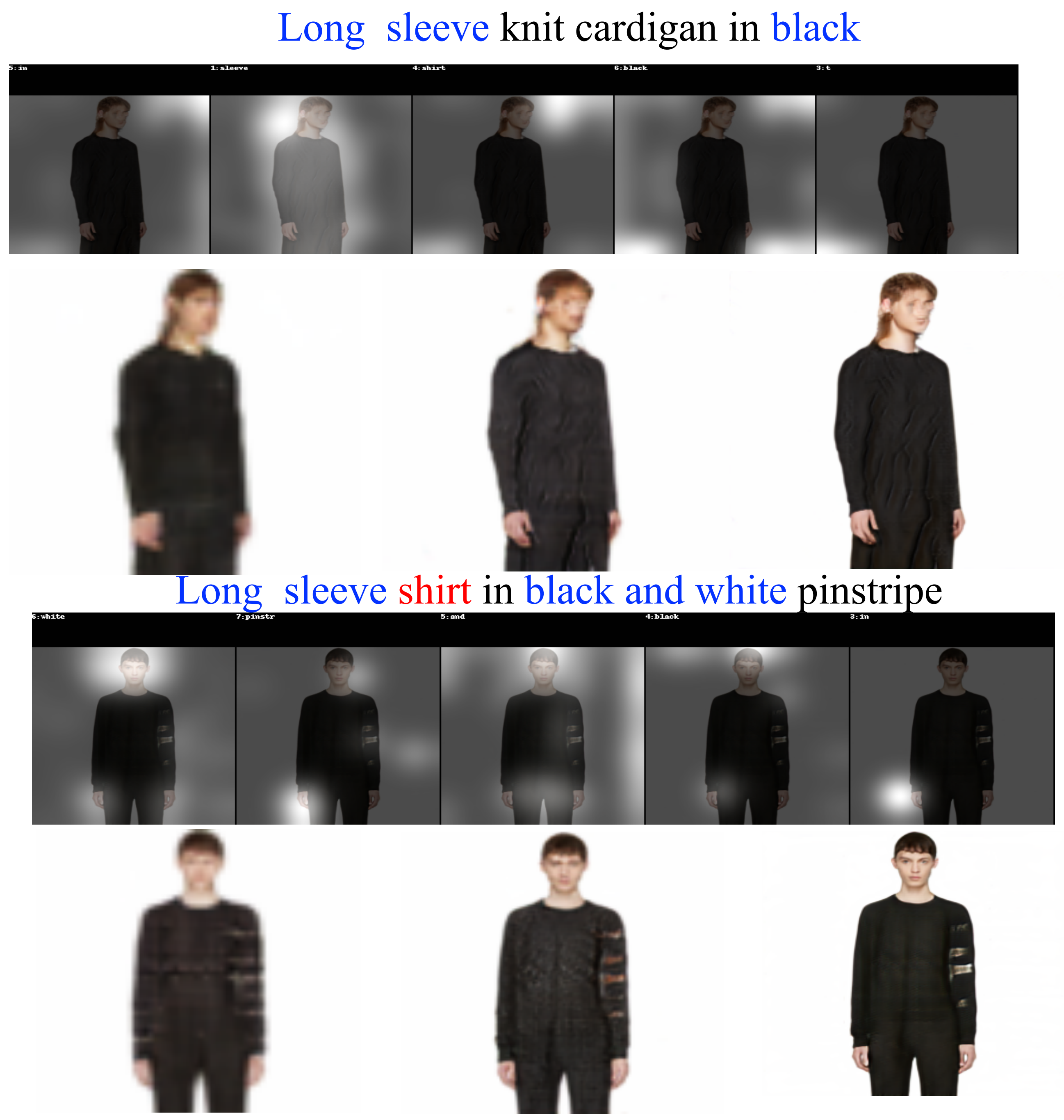
Samples at epoch 1



Samples at epoch 8







Long sleeve knit cardigan in black

Long sleeve shirt in black and white pinstripe

Conditional samples given text descriptions

Discussion

1. Both quantitative and qualitative results show the method is promising in obtaining reasonably high quality and realistic results.
2. Due to the computing resource and time limitation, this project could not evaluate the method using higher resolution images or training more epochs to gain better results.
3. One potential drawback we saw in the output results is the human faces are not as realistic as clothes.

Future

1. Train with more epochs
2. Test different hyperparameters
3. Add structural coherence for more realistic human faces and poses

References

1. T. Xu, P. Zhang, Q. Huang, H. Zhang, Z. Gan, X. Huang and X. He. AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks. In *CVPR*, 2018.