**Introduction**

- tremendous application: photo-editing, computer-aided design
- fashion gans: rapidly visualize and modify ideas, or share their ideas with others

**Difficulties**

- It is very difficult to capture details in original text description
- It is very difficult to train GAN to generate high-resolution photo-realistic images from text descriptions.

**Methods**

The whole architecture is follow AttnGAN[5]

- attentional generative network: draw different sub-regions of the image by focusing on words that are more relevant to sub-region being drawn
- deep attentional multimodal similarity model (DAMSM): compute the similarity between the generated image and the sentence
  - text encoder: bi-directional LSTM that extracts semantic vectors from text description
  - image encoder: Convolutional Neural Network (CNN) that maps images into semantic vectors

Loss Function: \( L = L_c + \lambda_{\text{DAMSM}} \), where \( L_c = \sum L_{c_i} \)

**Datasets**

- The FashionGan Challenge
  - items: quantity
    - number of images: 325,536
    - resolution: 1360 x 1360
    - categories: 48
    - pose: multiple
    - number of items: 78,850

**Results**

**DAMSM Training Results**

**AttentionGAN Training Results**

**More Results**

**Discussion**

**Summary**

- implement a conditional-gan, generate realistic fashion photo's based on text description
- use attention generative model to build relations between specific word and image region
- two-stage refinement product: high-resolution outputs (256 x 256)

Discuss limitations or future work:

1. high-resolution image (1280 x 1280)
2. multi-modal learning to transfer image to captions

**References**