IEEE International Conference on Automatic Face and Gesture Recognition

FG 2021

Spatially Constrained GAN for Face and Fashion Synthesis

Songyao Jiang[1], Hongfu Liu[2], Yue Wu[1] and Yun Fu[1]
[1]Northeastern University, Boston MA USA
[2]Brandeis University, Waltham MA, USA

Smile lab — Synergetic Media Learning Lab

Northeastern University

Problem Definition and Contribution

Goal: To provide spatial constraints to the image synthesis process and make it fully controllable, our goal is to decouple the image synthesis task into three dimensions (i.e., spatial, attribute and latent dimensions), which can be described as finding the mapping

$$G(z, c, s) \rightarrow y, \tag{1}$$

where z is the latent vector of size, s is the target segmentation, c is the target attribute, and y is the target image.

Motivations:

- Face and fashion synthesis are inherently one-to-many mapping from semantic segmentations to real images.
- Lacking spatial constraint, existing methods are not explicitly controllable in spatial dimension.

Key Contributions:

- We propose SCGAN that decouples the face and fashion synthesis task into three dimensions (spatial, attribute, and latent).
- A particularly designed generator extracts spatial information from segmentation, utilizes variations in random latent vectors and applies specified attributes. A segmentor network guides the generator with spatial constraints and improves model convergence.
- Extensive experiments on the CelebA and DeepFashion datasets demonstrate the effectiveness of SCGAN.

Problem Formulation

Main idea: We formulate the image synthesis task as a three-dimension mapping $G(z, c, s) \rightarrow y$, and train a generator network G to fit our desired mapping function. G takes three inputs which are a latent code z, an attribute label c, and a target segmentation map s to synthesize the desired image y.

We propose a segmentor network S to provide spatial constraints to G, and employ a discriminator network D with an auxiliary classifier D_c to guide G as well.

Objective Functions:

- Adversarial Loss:

$$\mathcal{L}_{adv} = L_{adv}^{real} + L_{adv}^{fake} + L_{gp}, \tag{2}$$

- Classification Loss:

$$\mathcal{L}_{cls}^{real} = \mathbb{E}_{x,c}\left[A_c(c, D_c(x))\right], \tag{3}$$

$$\mathcal{L}_{cls}^{fake} = \mathbb{E}_{z,c,s}\left[A_c(c, D_c(G(z, c, s)))\right], \tag{4}$$

where $A_c(\cdot, \cdot)$ computes the binary cross-entropy.

- Segmentation Loss:

$$\mathcal{L}_{seg}^{real} = \mathbb{E}_{x,s}[A_s(s, S(x)], \tag{5}$$

$$\mathcal{L}_{seg}^{fake} = \mathbb{E}_{z,c,s}\left[A_s(s, S(G(z, c, s)))\right], \tag{6}$$

where $A_s(\cdot, \cdot)$ computes the pixelwise cross-entropy.

Experiments & Results

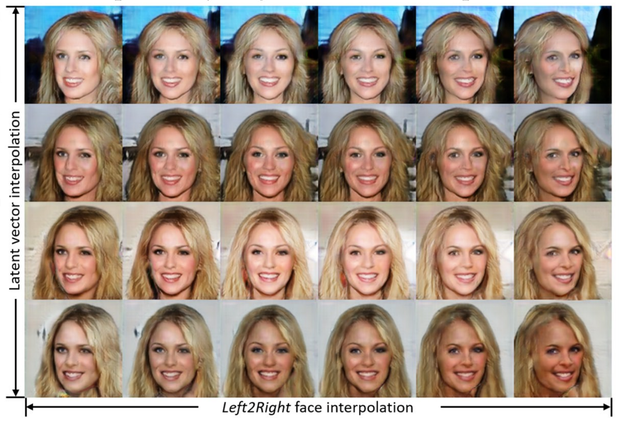
- We verify the effectiveness of SCGAN on a face dataset (CelebA) and a fashion dataset (DeepFashion) with both semantic segmentation and attribute label available.
- We show both visual and quantitative results compared with four representative methods, and present the spatial interpolation ability of SCGAN in face synthesis.

Quantitative Evaluation:

- We use four metrics (FID, pixel accuracy, and mean IoU) to evaluate SCGAN on CelebA and DeepFashion.

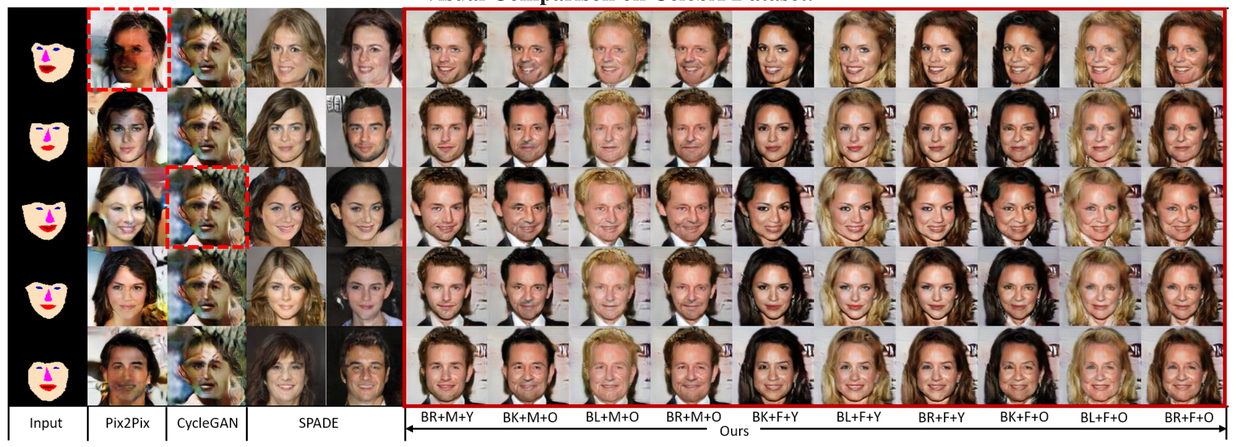
Methods	CelebA			DeepFashion		
	FID	mIoU	pAcc	FID	mIoU	pAcc
CycleGAN	N/A	N/A	N/A	30.1	63.26	82.21
Pix2Pix	20.4	78.71	98.05	24.4	65.41	82.91
SPADE	18.5	74.76	97.82	20.2	75.80	83.10
SCGAN	**10.2**	**79.11**	**98.95**	**19.8**	**77.20**	**83.23**

Samples of *Left2Right* and Latent Interpolation.

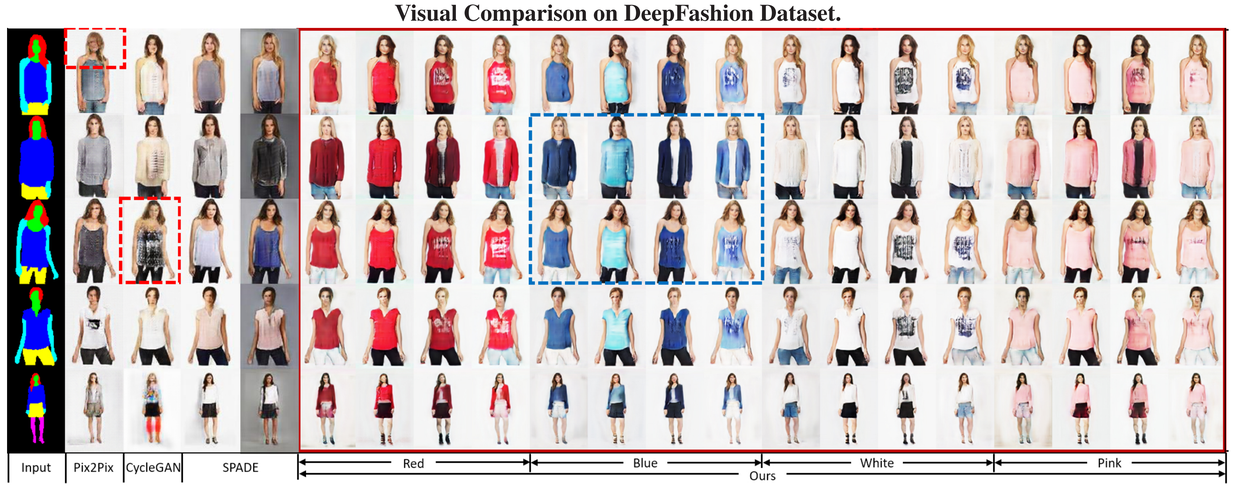


Qualitative Results:

Visual Comparison on CelebA Dataset.



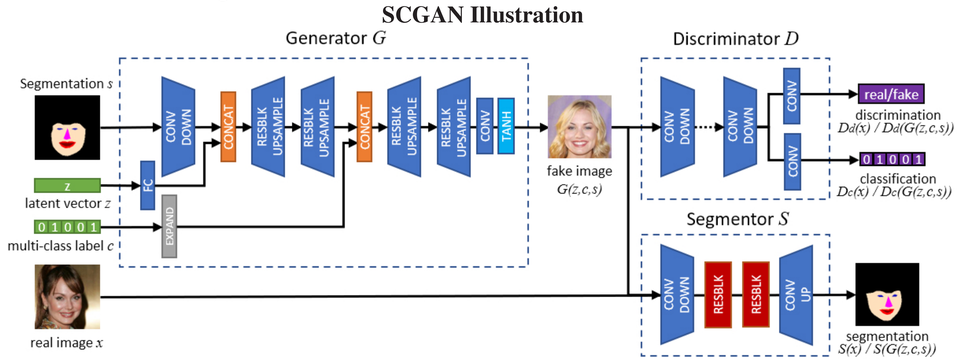
Visual Comparison on DeepFashion Dataset.



Method

Our proposed SCGAN consists of three networks as shown below, which are a generator network G, a discriminator network D, and a segmentor network S, as illustrated below

SCGAN Illustration



Scan QR Code

See training algorithm, code, more results and other details on our project webpage

Loss terms:

- $\mathcal{L}_{adv}$: Adversarial losses.
- $\mathcal{L}_{cls}^{real}$: Classification loss for real images.
- $\mathcal{L}_{cls}^{fake}$: Classification loss for synthesized fake images.
- $\mathcal{L}_{seg}^{real}$: Segmentation loss for real images.
- $\mathcal{L}_{seg}^{fake}$: Segmentation loss for fake images.

Overall objectives to optimize SCGAN:

$$\mathcal{L}_S = \mathcal{L}_{seg}^{real}, \tag{7}$$

$$\mathcal{L}_D = -\mathcal{L}_{adv} + \lambda_{cls}\mathcal{L}_{cls}^{real}, \tag{8}$$

$$\mathcal{L}_G = \mathcal{L}_{adv}^{fake} + \lambda_{cls}\mathcal{L}_{cls}^{fake} + \lambda_{seg}\mathcal{L}_{seg}^{fake}, \tag{9}$$