

Spatially Constrained GAN for Face and Fashion Synthesis

Songyao Jiang¹, Hongfu Liu², Yue Wu¹ and Yun Fu^{1,3}

¹Department of Electrical and Computer Engineering, Northeastern University, Boston MA, USA

²Michtom School of Computer Science, Brandeis University, Waltham MA, USA

³Khoury College of Computer Science, Northeastern University, Boston MA, USA

Abstract— Image synthesis has raised tremendous attention in both academic and industrial areas, especially for conditional and target-oriented image synthesis, such as criminal portrait and fashion design. The current studies have achieved encouraging results along this direction, but they mostly focus on class labels where spatial contents are randomly generated from latent vectors. Some recent studies have explored spatial constraints for generative models guided by semantic segmentation, but most of them are designed for scene generation and lack random variation. Such methods are not suitable for face or fashion image synthesis, where different images may share the same semantics. Different from all the current methods, we decouple the image synthesis task into three independent dimensions and propose a novel Spatially Constrained Generative Adversarial Network (SCGAN) to model it. SCGAN uses a simple yet effective way to decouple spatial constraints and attribute conditions from latent vectors, and treat them as additional controllable signals via a segmentor and a specially designed generator. Other unregulated contents are left to be generated from latent vectors. Experimentally, we provide both qualitative and quantitative results on CelebA and DeepFashion datasets to demonstrate that the proposed SCGAN is very effective in synthesizing spatially controllable and attribute-specific images with high visual quality and large variations. Our code is provided at <https://github.com/jackyjsy/SCGAN>.

I. INTRODUCTION

The success of Generative Adversarial Networks (GAN) [10] upsurges an increasing trend of realistic image synthesis [49], [47], [42], where a generator network produces artificial samples to mimic the real samples from a given dataset and a discriminator network attempts to distinguish between the real samples and artificial samples. These two networks are trained adversarially as two players in a game, and eventually, the two-player game will end up with the Nash Equilibrium. In such equilibrium, the generator is capable of mapping latent vectors from a simple distribution to real data samples from a complex distribution, while the discriminator can hardly distinguish the artificial samples from the real ones. GANs have been widely used in many applications such as natural language processing [48], [46], image super-resolution [23], [30], domain adaptation [14], [5], object detection [25], activity recognition [26], video prediction [34], face aging [29], semantic segmentation [33], face frontalization [44], [45], and image translations [15], [50], [17].

Beyond generating arbitrary images, conditional and target-oriented image generation is highly needed in various practical scenarios, such as criminal portraits based on victims' descriptions, clothing design with certain fashion elements, data augmentation, and artificial intelligence

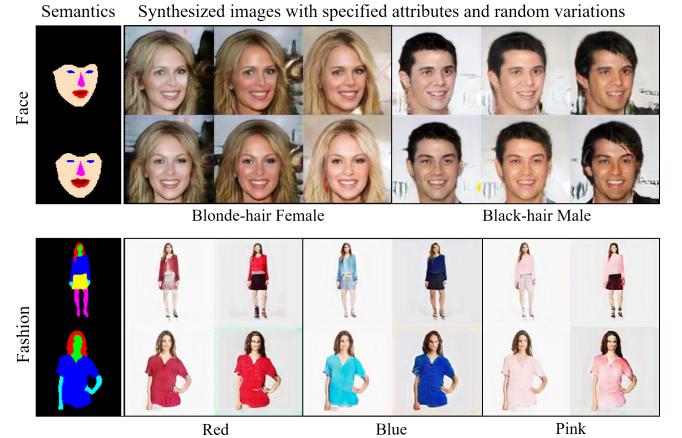


Fig. 1. SCGAN decouples the image synthesis task into three dimensions (*i.e.*, spatial, attribute and latent dimensions). SCGAN synthesizes face and fashion images guided by target semantic segmentations, specified attributes and achieves large variations on other unregulated components (*e.g.*, textures, skin colors, hair styles, fashion design, and color shades).

imagination. cGAN [35] first provided a way of conditional generation according to input class labels, which is further extended by [37] and [8] that additional classifiers are utilized to guide the image generation. They focus on available class labels as the condition where spatial contents are still randomly constructed from latent vectors. The edge details are usually blurred and the boundary information is difficult to preserve due to the lack of spatial constraints. Semantic-guided image synthesis has been recently explored in [38], [40] for scene image synthesis. Those methods use image-to-image translation networks to generate scene images from semantic segmentations. However, when applied to face and fashion image synthesis, those methods cannot provide much diversity with given semantic segmentations. In other words, they are deterministic and tend to synthesize fixed outputs with given input semantics. Some efforts such as SPADE [38] encode a style image to a style vector to obtain diverse outputs. Such design works well for scene images, however, our experiments reveal that such method does not provide a good diversity for face or fashion synthesis.

For face and fashion synthesis, inherently, there exists a one-to-many mapping from semantic segmentations to real images. Many distinct faces and clothes could share very similar semantics but retain diverse textures and attributes. This is a major reason why those image-to-image translation-based semantic-guided image synthesis methods are not suitable for face and fashion synthesis tasks. To solve the problem, we propose to decouple the face and fashion

synthesis tasks into three dimensions, which are spatial dimension, attribute dimension, and latent dimension, and make the first two dimensions explicitly controllable. The spatial configurations of generated images are regulated by input semantic segmentations, the attributes are specified by input attribute labels and the other uncontrolled components are automatically synthesized from input latent vectors.

We propose a Spatially Constrained Generative Adversarial Network (SCGAN) to learn the mapping of the three-dimension image synthesis. SCGAN consists of three networks, a generator network, a discriminator network with an auxiliary classifier, and a segmentor network, which are trained together adversarially. The generator is specially designed to take a semantic segmentation, a latent vector, and an attribute label as inputs step by step to synthesize a fake image. The discriminator network tries to distinguish between real images and generated images as well as classifying them into multi-label attributes. The discrimination and classification results guide the generator to synthesize realistic images with correct target attributes. The segmentor network attempts to estimate semantic segmentations on both real images and fake images to deliver estimated segmentations, which guides the generator in synthesizing spatially constrained images. With those networks, the proposed SCGAN generates realistic and diverse face and fashion images guided by input semantic segmentations and attribute labels, which enables many interesting applications such as interpolating between left faces and their faces, and generating intermediate faces from not smiling to smiling facial expression. Experimentally, we demonstrate the effectiveness and benefits of the spatial constraints by providing both qualitative and quantitative results on a face dataset CelebA [32] and a fashion dataset DeepFashion [31]. Here we highlight our major contributions as follows.

- We decouple the face and fashion synthesis task into three dimensions (*i.e.*, spatial, attribute, and latent) and propose a novel SCGAN to model it. Both spatial and attribute dimensions can be explicitly controllable.
- A generator network is particularly designed to extract spatial information from input segmentation, then concatenate a latent vector to provide variations and apply specified attributes. A segmentor network is introduced to guide the generator with spatial information and increases the model stability for convergence.
- Extensive experiments are conducted on the CelebA and DeepFashion datasets to demonstrate that the proposed SCGAN is effective in controlling spatial and attribute contents and can synthesize face and fashion images with large variations.

II. RELATED WORK

In recent years, deep generative models inspired by GAN enable computers to synthesize new samples based on the knowledge learned from given datasets. There have been many variations of GAN to improve the generating ability and stabilize adversarial training such as [39], [2], [1], [11], [28], [36], [4], [18], [20], [19]. In the meanwhile, many

researchers focused on developing target-oriented generative models instead of random generation. Conditional GAN [35] is the first attempt to input conditional labels into both generator and discriminator to achieve conditional image generation. Similarly, ACGAN [37] constructs an auxiliary classifier within the discriminator to output classification results and TripleGAN [8] introduces a classifier network as an extra player to the original two players setting. But all these studies focus on attribute-level conditions and neglect spatial conditions, which leads to the lack of spatial controllability in synthesized images.

People have been working on manipulating spatial contents of images via 3D morphable models since 1990s [3]. Recently, synthesizing spatially constrained images via a GAN-based network is first exploited using image-to-image translation methods, where input images can be regarded as spatial conditions in image translation. Pix2Pix [15] is the first to use an image as the conditional input and trains their networks with supervision from paired images. Then many researchers find out that paired training is unnecessary after introducing a cycle-consistency loss and propose several unpaired image translation methods [50], [21], [43], [27], [42], [16]. Based on those two-domain translation methods, StarGAN [6] proposes a multi-domain image translation network with an auxiliary classifier. Vid2Vid [41] further extends the image translation to a video translation, which enables many interesting applications such as synthesizing dance videos from skeleton videos. Human body pose landmarks are used as spatial constraints to guide generative networks and synthesize whole-body images in [12], [9].

Most recently, MaskGAN [24] utilizes facial attribute masks to enable interactive face image manipulation, SPADE [38] proposes a spatially-adaptive normalization to effectively generate high-resolution images based on given semantic segmentation with different learned styles, and LGGAN [40] further improves the ability of semantic-guided scene generation to synthesize small objects and detailed local textures. Style-guided image translation methods [7], [51] merge a base image with a style image to synthesize a new image that spatial configuration and attribute-level contents are inherited and not explicitly controllable. All the above methods have intrinsic assumptions of one-to-one mappings that they synthesize deterministic images without much variation. Therefore, those methods perform quite well for scene synthesis but are not suitable for conditional face or fashion generation which require large variations.

Different from all existing methods, SCGAN decouples the image generation into three dimensions via a simple and novel design of networks, and utilizes semantic segmentation as spatial constraints in a distinctive way. SCGAN takes a latent vector, attribute labels, and a semantic segmentation as inputs, explicitly controls spatial configurations and attribute contents, and generates target images with a large diversity.

III. METHODOLOGY

In this section, we first define our target problem and define the symbols used in our methodology. Then we

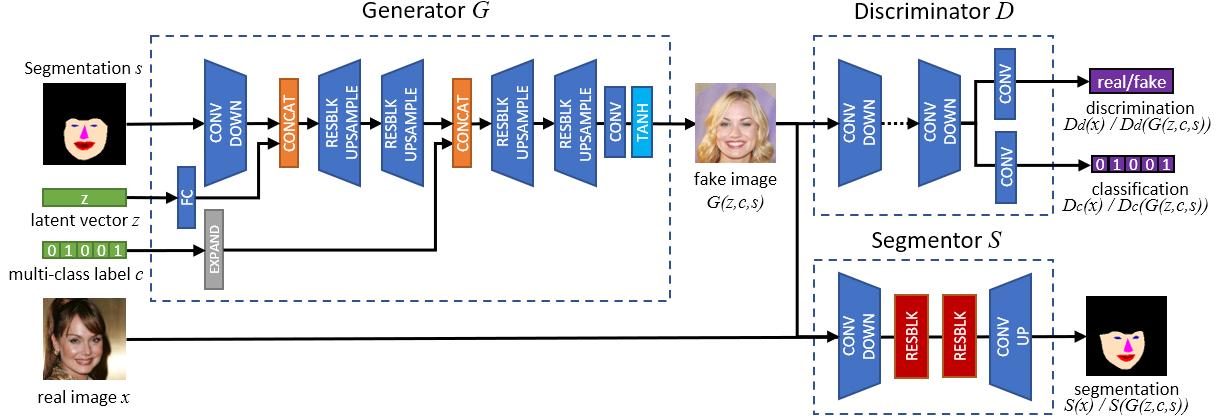


Fig. 2. SCGAN consists of a generator, a discriminator with an auxiliary classifier and a segmentor which are trained together. The generator is particularly designed that a semantic segmentation, a latent vector and an attribute label are input to the generator step by step to generate a fake image. The discriminator takes either a fake or real image as input and outputs a discrimination result and a classification result. The segmentor takes either fake or real image as input and outputs a segmentation result, and guides the generator to synthesize fake images which comply with the target segmentation.

introduce the framework structure of the proposed SCGAN. After that, all loss terms in the objective functions to optimize those networks are discussed in detail. Last, we provide a detailed training algorithm.

A. Problem Setting

Let $\mathbb{P}(x, c, s)$ denotes the joint distribution of the target joint dataset with attribute labels and geometric configuration, where x is a real image of size $(H \times W \times 3)$, c is its multi-attribute label of size $(1 \times n_c)$ with n_c as the number of attributes, and s is its semantic segmentation of size $(H \times W \times n_s)$ with n_s as the number of segmentation classes. Each pixel in s is represented by an one-hot vector with dimension n_s , which codes the semantic index of that pixel. Our goal can be described as finding the mapping $G(z, c, s) \rightarrow y$, where $G(\cdot, \cdot, \cdot)$ is the generating function, z is the latent vector of size $(1 \times n_z)$, and y is the conditionally generated image which complies with the target conditions c and s . Our target can be expressed as training a deep generator network to fit the target mapping function $G(z, c, s) \rightarrow y$, where the joint distribution $\mathbb{P}(y, c, s)$ is expected to follow the same distribution as $\mathbb{P}(x, c, s)$.

B. Spatially Constrained Generative Adversarial Networks

In this paper, we propose a generative model called Spatially Constrained Generative Adversarial Networks (SCGAN) to help training a generator network to fit the target mapping function $G(z, c, s) \rightarrow y$. Our proposed SCGAN consists of three networks shown in Figure 2, which are a generator network G , a discriminator network D , and a segmentor network S . Here we introduce each network individually in detail, define their objective functions, and provide a training algorithm to optimize these networks.

Generator Network. We utilize a generator network G to match our desired mapping function $G(z, c, s) \rightarrow y$. Our generator takes three inputs which are a latent code z , an attribute label c , and a target segmentation map s . As shown in Figure 2, these inputs are fed into the generator step by step in orders. First, the generator G takes s as input

and extracts spatial information contained in s by several downsampling convolutional layers. After that, the convolution result is concatenated with a dimensional expansion of z in channel dimension. After a few upsampling residual blocks (RESBLKUP), c is fed into the generator at last to guide the generator to generate attribute-specific images that contain basic image contents generated from s and z . This particular design of G decides the spatial configuration of the synthesized image according to the spatial constraints extracted from s . Then G forms the basic structure (*e.g.*, background, ambient lighting) of the generated image using the information coded in z . After that, G generates the attribute components specified by c .

Discriminator Network. To obtain realistic results which can hardly be distinguished from the real images, we employ a discriminator network D which forms a GAN framework with G . An auxiliary classifier is embedded in D to do a multi-class classification which provides attribute-level and domain-specific information back to G . D is defined as $D : x \rightarrow \{D_d(x), D_c(x)\}$, where $D_d(x)$ gives the discrimination results and $D_c(x)$ outputs the probabilities of x belonging to n_c attributes. D and G are two adversarial players in training, which eventually makes $\mathbb{P}(G(z, c, s), c)$ close to $\mathbb{P}(x, c)$.

Segmentor Network. We propose a segmentor network S to provide spatial constraints in conditional image generation. Let $S(\cdot)$ be the mapping function. S takes either real or generated image data as input and outputs the probabilities of pixel-wise semantic segmentation results of size $(H \times W \times n_s)$. S can be trained solely using x with its corresponding s . When training the other networks SCGAN, the weights in S can be fixed, and S can still provide the gradient information to G . Training S separately speeds up the model convergence and reduces the memory usage of the GPUs.

C. Objective Functions

Adversarial Loss. We adopt a conditional objective from Wasserstein GAN with gradient penalty [11]

$$\mathcal{L}_{adv} = L_{adv}^{real} + L_{adv}^{fake} + L_{gp}, \quad (1)$$

which can be rewritten as

$$\begin{aligned}\mathcal{L}_{adv} = & \mathbb{E}_x [D_d(x)] + \mathbb{E}_{z,c,s} [D_d(G(z, c, s))] \\ & + \lambda_{gp} \mathbb{E}_{\hat{x}} \left[(\|\nabla_{\hat{x}} D_d(\hat{x})\|_2 - 1)^2 \right],\end{aligned}\quad (2)$$

where $G(z, c, s)$ is the generated image conditioned on both attribute label c and segmentation s , λ_{gp} controls the weight of gradient penalty term, \hat{x} is the uniformly interpolated samples between a real image x and its corresponding fake image $G(z, c, s)$. During the training process, D and G act as two adversarial players that D tries to maximize this loss while G tries to minimize it.

Segmentation Loss acts as a spatial constraint to regulate the generator to comply with the spatial information defined by the input semantic segmentation. The proposed real segmentation loss to optimize the segmentor network S can be described as

$$\mathcal{L}_{seg}^{real} = \mathbb{E}_{x,s} [A_s(s, S(x))], \quad (3)$$

where $A_s(\cdot, \cdot)$ computes cross-entropy pixel-wisely by

$$A_s(a, b) = - \sum_{i=1}^H \sum_{j=1}^W \sum_{k=1}^{n_s} a_{i,j,k} \log b_{i,j,k}, \quad (4)$$

where a is the ground-truth segmentation and b is the estimated segmentation of a of size $(H \times W \times n_s)$. Taking a real image x as input, estimated segmentation $S(x)$ is compared with ground-truth segmentation s to optimize the segmentor S . When training together with the generator G , the segmentation loss term to optimize G is defined as

$$\mathcal{L}_{seg}^{fake} = \mathbb{E}_{z,c,s} [A_s(s, S(G(z, c, s)))], \quad (5)$$

where the estimated segmentation $S(G(z, c, s))$ is compared with the input segmentation s . By minimizing this loss term, the generator is forced to generate fake images which are consistent with the input semantic segmentations s .

Classification Loss. We embed an auxiliary multi-attribute classifier D_c which shares the weights with D_d in discriminator D except the output layer. The auxiliary classifier D_c takes an image as input and classify the image into independent probabilities of n_c attribute labels. The classification loss for real samples is defined as

$$\mathcal{L}_{cls}^{real} = \mathbb{E}_{x,c} [A_c(c, D_c(x))], \quad (6)$$

where (x, c) is a pair of real image with its attribute label, $A_c(\cdot, \cdot)$ computes a multi-attribute binary cross-entropy loss by $A_c(a, b) = - \sum_k a_k \log(b_k)$ with a, b being two vectors of identical size $(1 \times n_c)$. Accordingly, we have the classification loss for fake samples by

$$\mathcal{L}_{cls}^{fake} = \mathbb{E}_{z,c,s} [A_c(c, D_c(G(z, c, s)))], \quad (7)$$

which takes the fake image $G(z, c, s)$ as input and guides G to generate attribute-specific images according to the classification information learned from real samples.

Overall Objectives to optimize S , D and G in SCGAN can be represented as

$$\mathcal{L}_S = \mathcal{L}_{seg}^{real}, \quad (8)$$

Algorithm 1: Training SCGAN, where $\lambda_{cls} = 5$,

$\lambda_{seg} = 1$, $\lambda_{gp} = 10$, $n_{repeat} = 5$ and $m = 16$.

```

1 Initialize three network parameters  $\theta_G$   $\theta_D$ ,  $\theta_S$ ;
2 while  $\theta_G$  has not converged do
3   for  $n = 1, \dots, n_{repeat}$  do
4     Sample a batch of latent vectors
5        $\{z^i\}_{i=1}^m \sim \mathcal{N}(0, 1)$ ;
6     Sample a batch of  $\{x^i, c^i, s^i\}_{i=1}^m$  from
7        $\mathbb{P}_{data}(x, c, s)$ ;
8     Sample a batch of numbers
9        $\{\epsilon^i\}_{i=1}^m \sim \mathcal{U}(0, 1)$ ;
10     $\{s_t^i\}_{i=1}^m \leftarrow \text{shuffle}(\{s^i\}_{i=1}^m)$ ;
11    for  $i = 1, \dots, m$  do
12       $\tilde{x}^i \leftarrow G(z^i, c^i, s_t^i)$ ;
13       $\hat{x}^i \leftarrow \epsilon^i x^i + (1 - \epsilon^i) \tilde{x}^i$ ;
14       $\mathcal{L}_{adv}^i \leftarrow D_d(\hat{x}^i) - D_d(x^i)$ 
15         $+ \lambda_{gp} (\|\nabla_{\hat{x}} D_d(\hat{x}^i)\|_2 - 1)^2$ ;
16       $\mathcal{L}_{cls}^{real,i} \leftarrow A_c(c^i, D_c(x^i))$ ;
17       $\mathcal{L}_{seg}^{real,i} \leftarrow A_s(s^i, S(x^i))$ ;
18    end
19    Update  $D$  by descending its gradient:
20       $\nabla_{\theta_D} \frac{1}{m} \sum_i^m \mathcal{L}_{adv}^i + \lambda_{cls} \mathcal{L}_{cls}^{real,i}$ ;
21    Update  $S$  by descending its gradient:
22       $\nabla_{\theta_S} \frac{1}{m} \sum_i^m \mathcal{L}_{seg}^{real,i}$ ;
23  end
24  for  $i = 1, \dots, m$  do
25     $\tilde{x}^i \leftarrow G(z^i, c^i, s_t^i)$ ;
26     $\mathcal{L}_{adv}^i \leftarrow D_d(\tilde{x}^i)$ ;
27     $\mathcal{L}_{cls}^{fake,i} \leftarrow A_c(c^i, D_c(\tilde{x}^i))$ ;
28     $\mathcal{L}_{seg}^{fake,i} \leftarrow A_s(s_t^i, S(\tilde{x}^i))$ ;
29  end
30  Update  $G$  by descending its gradient:
31     $\nabla_{\theta_G} \frac{1}{m} \sum_i^m (\mathcal{L}_{adv}^i + \lambda_{cls} \mathcal{L}_{cls}^{fake,i} + \lambda_{seg} \mathcal{L}_{seg}^{fake,i})$ ;
32 end
```

$$\mathcal{L}_D = -\mathcal{L}_{adv} + \lambda_{cls} \mathcal{L}_{cls}^{real}, \quad (9)$$

$$\mathcal{L}_G = \mathcal{L}_{adv}^{fake} + \lambda_{cls} \mathcal{L}_{cls}^{fake} + \lambda_{seg} \mathcal{L}_{seg}^{fake}, \quad (10)$$

where \mathcal{L}_S , \mathcal{L}_D and \mathcal{L}_G are objective functions to optimize S , D and G . λ_{seg} and λ_{cls} are hyper-parameters which control the relative importance of \mathcal{L}_{seg} and \mathcal{L}_{cls} compared to \mathcal{L}_{adv} .

D. Training Algorithm

Let θ_G , θ_D and θ_S be the parameters of networks G , D and S , respectively. Our objective is to find a converged θ_G with minimized \mathcal{L}_G . When training the proposed SCGAN, a batch of latent vectors are sampled from a Gaussian distribution $\mathcal{N}(0, 1)$ denoted as $\{z^i\}_{i=1}^m$, where m is the batch size. A batch of x with its ground-truth s and c are randomly sampled from the joint distribution $\mathbb{P}_{data}(x, c, s)$ of the target dataset, denoted as $\{x^i, c^i, s^i\}_{i=1}^m$. When selecting target

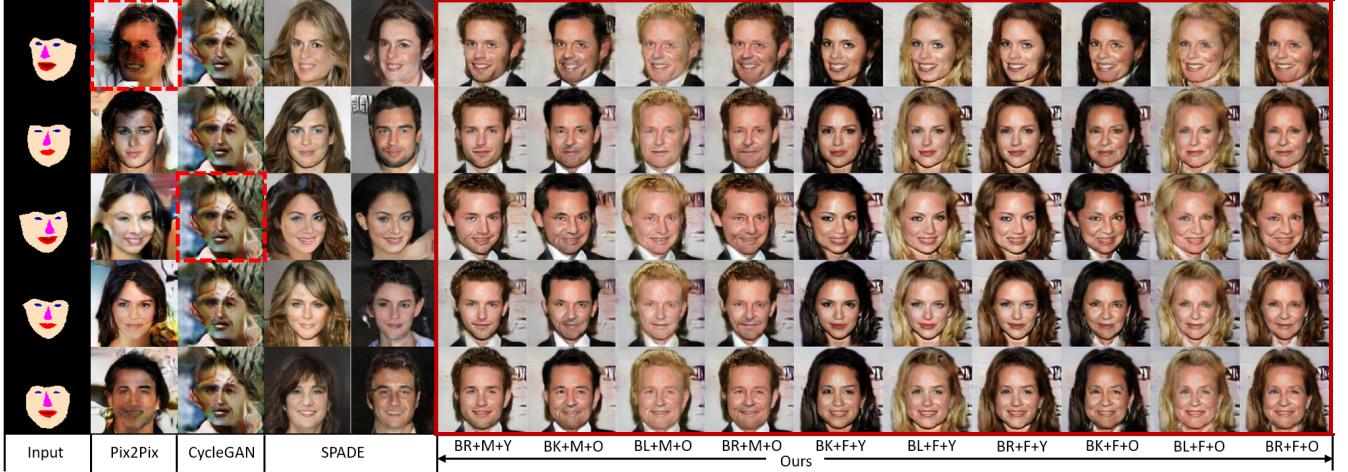


Fig. 3. Comparison results on CelebA dataset. Our results are shown in the solid red rectangle. Failure cases of the compared methods are highlighted by the dashed red rectangle. (Abbrev.: BL=Blond Hair, BR=Brown Hair, BK=Black hair, M=Male, F=Female, Y=Young, O=Old.)

semantic segmentation for $\{x^i\}_{i=1}^m$, $\{s^i\}_{i=1}^m$ are randomly shuffled to obtain a batch of target segmentations $\{s_t^i\}_{i=1}^m$ to be input to G . Details can be found in Algorithm 1.

IV. EXPERIMENT

In this section, we verify the effectiveness of SCGAN on a face dataset and a fashion dataset with both semantic segmentation and attribute label. We show both visual and quantitative results compared with four representative methods, present the spatial interpolation ability of our model in terms of face synthesis.

A. Datasets

Large-scale CelebFaces Attributes (CelebA) dataset [32] contains 202,599 face images of celebrities with 40 binary attribute labels and 5-point facial landmarks. We use the aligned version of face images and select 5 attributes including black hair, blond hair, brown hair, gender, and age in our experiment. This dataset doesn't provide any ground-truth semantic segmentation for the face images. To obtain the semantic segmentation, we apply Dlib [22] landmarks detector to extract 68-point facial landmarks from the faces images, which separate facial attributes into six different regions. By filling those regions with corresponding semantic index pixel-wisely, semantic segmentations are created.

Large-scale Fashion (DeepFashion) dataset [31], [52] is a large-scale clothing database, which contains over 800,000 diverse fashion images ranging from well-posed shop images to unconstrained photos from customers. In our experiment, we use one of the subsets particularly designed for the fashion synthesis task, which selects 78,979 clothing images from the In-shop Clothes Benchmark associated with their attribute labels, captions, and semantic segmentations. We use the 18-class color attributes and the provided semantic segmentation in our experiment.

B. Compared Methods

Pix2Pix [15] and CycleGAN [50] are two popular image-to-image translation method, which can take semantic seg-

mentation as input and synthesize realistic images. Pix2Pix requires paired images while CycleGAN is trained in an unpaired way. We also compare our method with a most recent state-of-the-art method named SPADE [38] which can generate images given semantic segmentations following the styles/modalities of input images. In our experiment, we use the official implementation released by the authors, train their model on our target datasets, and try our best to tune the parameters to deliver good results.

C. Spatially Constrained Face Synthesis

We first provide comparison results on CelebA in Figure 3. The input segmentations are shown in the leftmost column, and the results of Pix2Pix and CycleGAN are shown in the next two columns. The visual quality generated by Pix2Pix is low, and CycleGAN suffers a mode collapse issue that their model only gives a single output no matter the input segmentation. One possible reason is that translating facial segmentation to realistic faces is essentially a one-to-many translation, however, those two image-to-image translation methods both assume a one-to-one mapping between input and target domains. Especially for CycleGAN, their cycle-consistency loss which seeks to maintain the contents during a cycle translating forward and backward tends to enforce the one-to-one mapping. When a face image is translated into its semantic segmentation, it is barely possible to translate it back to the original face due to the information lost in the many-to-one translation. The results of SPADE are presented in Column 4 and 5 in Figure 3. SPADE can generate diverse faces given fixed segmentation as inputs, but the attributes of the generated faces are randomized despite providing “style images” to the encoder. Since face images are similar to each other in structures, it violates the style-based assumption of SPADE. Our proposed SCGAN could always produce reliable and high-quality results. It is worth noting that inputting randomly sampled latent vectors can result in diverse images with different backgrounds and details in segmentation-to-image synthesis. Due to the high-



Fig. 4. *NoSmile2Smile* facial expression interpolations. Each row shows a group of interpolated results between a not smiling face and a smiling face with a specific attribute label and a fixed latent vector.

TABLE I

QUANTITATIVE EVALUATION ON CELEBA AND DEEPFASHION DATASET USING FRÉCHET INCEPTION DISTANCE (FID), MEAN IOU (mIoU) AND PIXEL ACCURACY (pAcc). N/A INDICATES MODE COLLAPSE

Methods	CelebA			DeepFashion		
	FID	mIoU	pAcc	FID	mIoU	pAcc
CycleGAN [50]	N/A	N/A	N/A	30.1	63.26	82.21
Pix2Pix [15]	20.4	78.71	98.05	24.4	65.41	82.91
SPADE [38]	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

frequency signal from boundaries of attributes in semantic segmentation, our SCGAN could produce a large number of sharp details which makes the results more realistic compared to all the other methods. In summary, SCGAN enjoys superiority in terms of diverse variations, controllability, and realistic high-quality results over the other methods.

D. Interpolation Abilities

Beyond face synthesis, our proposed SCGAN can control the face orientation and facial expressions of the synthesized faces by feeding corresponding semantic segmentations as guidance. To synthesize faces of every intermediate state between two facial orientations and expressions, corresponding semantic segmentations of those intermediate states are needed. It is difficult to obtain such intermediate segmentations that numeric interpolation between two segmentations only results in a fade-in and fade-out effect. Instead, we interpolate every intermediate state on x - y coordinates of facial landmarks instead of segmentation domain. We then construct semantic segmentations from those landmarks to obtain spatial-varying semantic segmentation. As shown in Figure 4 and 5, SCGAN generates intermediate faces from not smiling face to smiling face (*NoSmile2Smile*) and from left-side face to right-side face (*Left2Right*). Interpolations on latent vectors are also shown in Figure 5.

E. Spatially Constrained Fashion Synthesis

Comparison results on the DeepFashion dataset presented in Figure 6 also demonstrate the advantages of our pro-



Fig. 5. Two-dimension interpolation results in latent space and between *Left2Right* faces. Each column presents the results of interpolated latent vectors, and each row shows the interpolation results on facial orientations.

posed SCGAN over the other methods. Similar to Figure 3, the input segmentation, results of Pix2Pix, CycleGAN and SPADE are shown in the left five columns. The images in the large solid red rectangle are our results from SCGAN with both semantic segmentation and attribute labels and latent vector as inputs. Different from the results on CelebA dataset, image-to-image translation methods are capable of producing acceptable results on the DeepFashion dataset, because the intrinsic one-to-many property in the DeepFashion dataset is not as strong as in the CelebA dataset. In the DeepFashion dataset, the ability of shape preserving becomes more important than general visual discrimination. Their results also lack attribute-level controllability and variations on fashion detail as our results highlighted by the dashed blue rectangle. With our semantic segmentation as the spatial constraints, SCGAN can generate fashion images controlled by the input color labels and semantic segmentation, while the input latent vectors encode variant fashion style (e.g., cardigans, T-shirts), diverse pants and shoes, and different color shades and saturation (e.g., dark blue, light blue).

F. Quantitative Evaluation

To quantitatively evaluate the effectiveness of spatially constrained image generation, we use Fréchet Inception Distance (FID) [13] to evaluate the fidelity of the generated images. FID measures the distance between real and synthesized data in their Inception embeddings. We also adopt metrics of mean IoU (intersection over union) and pixel accuracy to examine the spatial consistency between the input semantic segmentation and the generated images from the generator, which are commonly used when evaluating segmentation algorithms. We run the experiment for five times and report the averaged results compared with the image-to-image translation methods of CycleGAN, Pix2Pix and SPADE. As shown in Table I, our SCGAN achieves the best performance on both CelebA and DeepFashion datasets. Our method is capable of generating realistic images with diversity as well as make those images comply with the input semantic segmentations accurately.

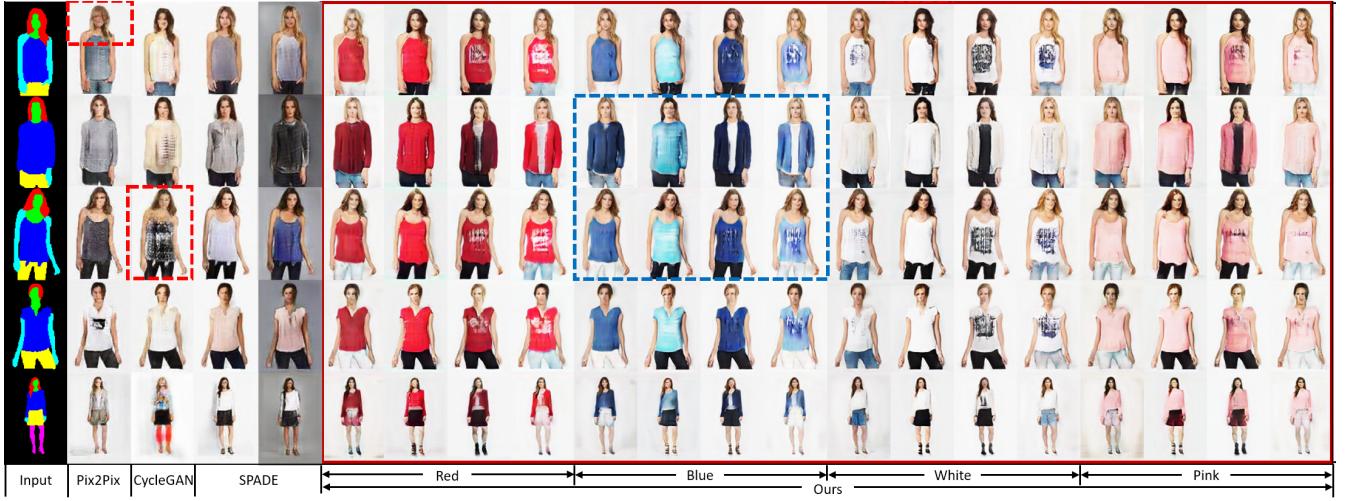


Fig. 6. Comparison with Pix2Pix, CycleGAN and SPADE on DeepFashion dataset. Their failure cases are highlighted in the dashed red rectangle, while the dashed blue rectangle highlights the representative diverse results generated by our proposed SCGAN.

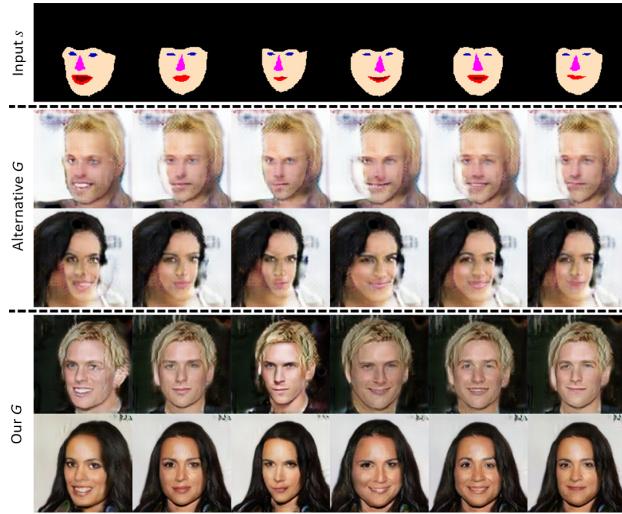


Fig. 7. An ablation study on generator configurations. Our proposed generator structure solves the foreground-background mismatch problem suffered by the alternative generator which inputs all the conditions at once.

G. Ablation Study on Generator Configuration

As described in Section III-B, the generator of our proposed SCGAN takes three inputs, a semantic segmentation, a latent vector, and an attribute label step by step in order that the contents in the synthesized image should be decoupled well to be controlled by those inputs. Otherwise, those inputs may conflict with each other and fail to generate the desired results. To demonstrate that, we conduct an ablation study to compare with an alternative generator that takes all the inputs and concatenates them together at the same time. We refer to this variant of generator network as the alternative G . As shown in Figure 7, severe foreground-background mismatches happen in the results of alternative G that the facial components regulated by the input segmentation cannot be merged correctly with the skin color or hairstyle determined by the latent vector. Our particularly designed generator could successfully decouple the contents

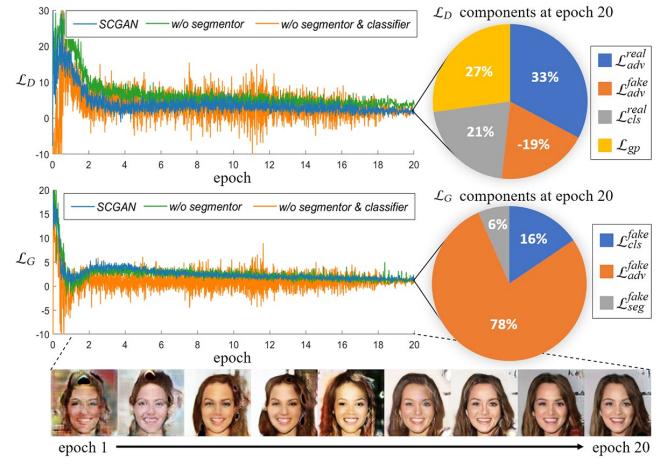


Fig. 8. An ablation study on model convergence. Losses during training are plotted together and the intermediate generated samples are shown.

of synthesized images into controllable inputs and generate variations on other unregulated contents.

H. Ablation Study on Model Convergence

Our proposed SCGAN converges fast and stably due to the introduction of the segmentor and the auxiliary classifier. We conduct an ablation study on model convergence by removing segmentor and auxiliary classifier. Figure 8 shows the losses of generator and discriminator during the training process on *CelebA* dataset. The blue plots are the losses of the proposed SCGAN. Green plots are the losses after removing the segmentor network. The orange plots show the losses after removing both the segmentor network and the embedded auxiliary classifier, while all the other things such as model architecture and hyper-parameters are kept unchanged. As observed from this figure, the training process of our SCGAN is much more stable with less vibration on losses. The convergence of SCGAN happens faster and its final loss is smaller than the other two ablation experiments.

The bottom part in Figure 8 shows the intermediate generated samples that improve gradually as the model converges.

V. CONCLUSIONS

In this paper, we proposed SCGAN to introduce spatial constraints to face and fashion image synthesis. Extensive experiments compared with other popular generative models on CelebA face dataset and DeepFashion datasets demonstrated that the proposed SCGAN was capable of controlling spatial contents, specifying attributes, and generating diversified images. We particularly designed the generator to take semantic segmentations, latent vectors, and attribute labels step by step to solve the foreground-background mismatch problem. In summary, our method is a simple yet effective variant of the GAN, which could be easily adapted to recent high-resolution GAN-based image generation models.

REFERENCES

- [1] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In *ICML*, 2017.
- [2] D. Berthelot, T. Schumm, and L. Metz. Began: Boundary equilibrium generative adversarial networks. *arXiv preprint arXiv:1703.10717*, 2017.
- [3] V. Blanz and T. Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, 1999.
- [4] A. Bora, E. Price, and A. G. Dimakis. Ambigantgan: Generative models from lossy measurements. In *ICLR*, 2018.
- [5] K. Bousmalis, N. Silberman, D. Dohan, D. Erhan, and D. Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *CVPR*, 2017.
- [6] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *CVPR*, 2018.
- [7] Y. Choi, Y. Uh, J. Yoo, and J.-W. Ha. Stargan v2: Diverse image synthesis for multiple domains. In *CVPR*, 2020.
- [8] L. Chongxuan, T. Xu, J. Zhu, and B. Zhang. Triple generative adversarial nets. In *NeurIPS*, 2017.
- [9] H. Dong, X. Liang, K. Gong, H. Lai, J. Zhu, and J. Yin. Soft-gated warping-gan for pose-guided person image synthesis. In *NeurIPS*, 2018.
- [10] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NeurIPS*, 2014.
- [11] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville. Improved training of wasserstein gans. In *NeurIPS*, 2017.
- [12] K. Hamada, K. Tachibana, T. Li, H. Honda, and Y. Uchida. Full-body high-resolution anime generation with progressive structure-conditioned generative adversarial networks. In *ECCV*, 2018.
- [13] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, 2017.
- [14] J. Hoffman, D. Wang, F. Yu, and T. Darrell. Fcns in the wild: Pixel-level adversarial and constraint-based adaptation. *arXiv preprint arXiv:1612.02649*, 2016.
- [15] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017.
- [16] S. Jiang, Z. Tao, and Y. Fu. Segmentation guided image-to-image translation with adversarial networks. In *FG*, 2019.
- [17] S. Jiang, Z. Tao, and Y. Fu. Geometrically editable face image translation with adversarial networks. *TIP*, 2021.
- [18] T. Karras, T. Aila, S. Laine, and J. Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *ICLR*, 2018.
- [19] T. Karras, M. Aittala, S. Laine, E. Härkönen, J. Hellsten, J. Lehtinen, and T. Aila. Alias-free generative adversarial networks. *arXiv preprint arXiv:2106.12423*, 2021.
- [20] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019.
- [21] T. Kim, M. Cha, H. Kim, J. K. Lee, and J. Kim. Learning to discover cross-domain relations with generative adversarial networks. In *ICML*, 2017.
- [22] D. E. King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 2009.
- [23] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. P. Aitken, A. Tejani, J. Totz, Z. Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*, 2017.
- [24] C.-H. Lee, Z. Liu, L. Wu, and P. Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *CVPR*, 2020.
- [25] J. Li, X. Liang, Y. Wei, T. Xu, J. Feng, and S. Yan. Perceptual generative adversarial networks for small object detection. In *CVPR*, 2017.
- [26] X. Li, Y. Zhang, J. Zhang, Y. Chen, H. Li, I. Marsic, and R. S. Burd. Region-based activity recognition using conditional gan. In *ACMMM*, 2017.
- [27] M.-Y. Liu, T. Breuel, and J. Kautz. Unsupervised image-to-image translation networks. In *NeurIPS*, 2017.
- [28] M.-Y. Liu and O. Tuzel. Coupled generative adversarial networks. In *NeurIPS*, 2016.
- [29] S. Liu, Y. Sun, D. Zhu, R. Bao, W. Wang, X. Shu, and S. Yan. Face aging with contextual generative adversarial nets. In *ACMMM*, 2017.
- [30] W. Liu, X. Liu, H. Ma, and P. Cheng. Beyond human-level license plate super-resolution with progressive vehicle search and domain priori gan. In *ACMMM*, 2017.
- [31] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *CVPR*, 2016.
- [32] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *ICCV*, 2015.
- [33] P. Luc, C. Couprie, S. Chintala, and J. Verbeek. Semantic segmentation using adversarial networks. In *NeurIPS Workshop on Adversarial Training*, 2016.
- [34] M. Mathieu, C. Couprie, and Y. LeCun. Deep multi-scale video prediction beyond mean square error. In *ICLR*, 2016.
- [35] M. Mirza and S. Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.
- [36] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida. Spectral normalization for generative adversarial networks. In *ICLR*, 2018.
- [37] A. Odena, C. Olah, and J. Shlens. Conditional image synthesis with auxiliary classifier gans. In *ICML*, 2017.
- [38] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu. Semantic image synthesis with spatially-adaptive normalization. In *CVPR*, 2019.
- [39] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.
- [40] H. Tang, D. Xu, Y. Yan, P. H. Torr, and N. Sebe. Local class-specific and global image-level generative adversarial networks for semantic-guided scene generation. In *CVPR*, 2020.
- [41] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, G. Liu, A. Tao, J. Kautz, and B. Catanzaro. Video-to-video synthesis. In *NeurIPS*, 2018.
- [42] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *CVPR*, 2018.
- [43] Z. Yi, H. R. Zhang, P. Tan, and M. Gong. Dualgan: Unsupervised dual learning for image-to-image translation. In *ICCV*, 2017.
- [44] Y. Yin, S. Jiang, J. P. Robinson, and Y. Fu. Dual-attention gan for large-pose face frontalization. In *FG 2020*, 2020.
- [45] Y. Yin, J. P. Robinson, S. Jiang, Y. Bai, C. Qin, and Y. Fu. Superfront: From low-resolution to high-resolution frontal face synthesis. In *ACMMM*, 2021.
- [46] L. Yu, W. Zhang, J. Wang, and Y. Yu. Seqgan: Sequence generative adversarial nets with policy gradient. In *AAAI*, 2017.
- [47] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *ICCV*, 2017.
- [48] Y. Zhang, Z. Gan, and L. Carin. Generating text via adversarial training. In *NeurIPS Workshop on Adversarial Training*, 2016.
- [49] Y. Zhao, B. Deng, J. Huang, H. Lu, and X.-S. Hua. Stylized adversarial autoencoder for image generation. In *ACMMM*, 2017.
- [50] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017.
- [51] P. Zhu, R. Abdal, Y. Qin, and P. Wonka. Sean: Image synthesis with semantic region-adaptive normalization. In *CVPR*, 2020.
- [52] S. Zhu, S. Fidler, R. Urtasun, D. Lin, and C. C. Loy. Be your own prada: Fashion synthesis with structural coherence. In *ICCV*, 2017.