Spatial Fusion GAN for Image Synthesis

Fangneng Zhan
Nanyang Technological University
50 Nanyang Avenue, Singapore 639798
fnzhan@ntu.edu.sg

Hongyuan Zhu
Institute for Infocomm, A*STAR, Singapore
1 Fusionopolis Way, Singapore 138632
zhuh@i2r.a-star.edu.sg

Shijian Lu
Nanyang Technological University
50 Nanyang Avenue, Singapore 639798
shijian.lu@ntu.edu.sg

Abstract

Recent advances in generative adversarial networks (GANs) have shown great potentials in realistic image synthesis whereas most existing works address synthesis realism in either appearance space or geometry space but few in both. This paper presents an innovative Spatial Fusion GAN (SF-GAN) that combines a geometry synthesizer and an appearance synthesizer to achieve synthesis realism in both geometry and appearance spaces. The geometry synthesizer learns contextual geometries of background images and transforms and places foreground objects into the background images unanimously. The appearance synthesizer adjusts the color, brightness and styles of the foreground objects and embeds them into background images harmoniously, where a guided filter is introduced for detail preserving. The two synthesizers are inter-connected as mutual references which can be trained end-to-end without supervision. The SF-GAN has been evaluated in two tasks: (1) realistic scene text image synthesis for training better recognition models; (2) glass and hat wearing for realistic matching glasses and hats with real portraits. Qualitative and quantitative comparisons with the state-of-the-art demonstrate the superiority of the proposed SF-GAN.

1. Introduction

With the advances of deep neural networks (DNNs), image synthesis has been attracting increasing attention as a means of generating novel images and creating annotated images for training DNN models, where the latter has great potentials to replace the traditional manual annotation which is usually costly, time-consuming and unscalable. The fast development of generative adversarial networks (GANs) [9] in recent years opens a new door of automated image synthesis as GANs are capable of generating realistic images by concurrently implementing a generator and discriminator. Three typical approaches have been explored for GAN-based image synthesis, namely, direct image generation [27, 33, 1], image-to-image translation [55, 16, 22, 14] and image composition [21, 2].

Figure 1. The proposed SF-GAN is capable of synthesizing realistic images concurrently in geometric and appearance spaces. Rows 1 and 2 show a few synthesized scene text images and row 3 shows a few hat-wearing and glass-wearing images where the foreground texts, glasses and hats as highlighted by red-color boxes are composed with the background scene and face images harmoniously.

On the other hand, most existing GANs were designed to achieve synthesis realism either from geometry space or appearance space but few in both. Consequentially, most GAN-synthesized images have little contribution (many even harmful) when they are used in training deep network models. In particular, direct image generation still faces difficulties in generating high-resolution images due to the limited network capacity. GAN-based image composition is capable of producing high-resolution images [21, 2] by placing foreground objects into background images. But most GAN-based image composition techniques focus on geometric realism only (e.g. object alignment with con-

textual background) which often produce various artifacts due to appearance conflicts between the foreground objects and the background images. As a comparison, GAN-based image-to-image translation aims for appearance realism by learning the style of images of the target domain whereas the geometric realism is largely ignored.

We propose an innovative Spatial Fusion GAN (SF-GAN) that achieves synthesis realism in both geometry and appearance spaces concurrently, a very challenging task in image synthesis due to a wide spectrum of conflicts between the foreground objects and background images with respect to relative scaling, spatial alignment, appearance style, etc. The SF-GAN address these challenges by designing a geometry synthesizer and an appearance synthesizer. The geometry synthesizer learns the local geometry of background images with which the foreground objects can be transformed and placed into the background images unanimously. A discriminator is employed to train a spatial transformation network, targeting to produce transformed images that can mislead the discriminator. The appearance synthesizer learns to adjust the color, brightness and styles of the foreground objects for proper matching with the background images with minimum conflicts. A guided filter is introduced to compensate the detail loss that happens in most appearance-transfer GANs. The geometry synthesizer and appearance synthesizer are inter-connected as mutual references which can be trained end-to-end with little supervision.

The contributions of this work are threefold. First, it designs an innovative SF-GAN, an end-to-end trainable network that concurrently achieves synthesis realism in both geometry and appearance spaces. To the best of our knowledge, this is the first GAN that can achieve synthesis realism in geometry and appearance spaces concurrently. Second, it designs a fusion network that introduces guided filters for detail preserving for appearance realism, whereas most image-to-image-translation GANs tend to lose details while performing appearance transfer. Third, it investigates and demonstrates the effectiveness of GAN-synthesized images in training deep recognition models, a very important issue that was largely neglected in most existing GANs (except a few GANs for domain adaptation [14, 16, 22, 55]).

2. Related Work

2.1. Image Synthesis

Realistic image synthesis has been studied for years, from synthesis of single objects [29, 30, 40] to generation of full-scene images [8, 34]. Among different image synthesis approaches, image composition has been explored extensively which synthesizes new images by placing foreground objects into some existing background image. The target is to achieve composition realism by controlling the object size, orientation, and blending between foreground objects and background images. For example, [10, 17, 50, 51] investigate synthesis of scene text images for training better scene text detection [47] and recognition models [49]. They achieve the synthesis realism by controlling a series of parameters such as text locations within the background image, geometric transformation of the foreground texts, blending between the foreground text and background image, etc. Other image composition systems have also been reported for DNN training [7], composition harmonization [26, 42], image inpainting [54], etc.

Optimal image blending is critical for good appearance consistency between the foreground object and background image as well as minimal visual artifacts within the synthesized images. One straightforward way is to apply dense image matching at pixel level so that only the corresponding pixels are copied and pasted, but this approach does not work well when the foreground object and background image have very different appearance. An alternative way is to make the transition as smooth as possible so that artifacts can be hidden/removed within the composed images, e.g. alpha blending [43], but this approach tends to blur fine details in the foreground object and background images. In addition, gradient-based techniques such as Poisson blending [31] can edit the image gradient and adjust the inconsistency in color and illumination to achieve seamlessly blending.

Most existing image synthesis techniques aim for geometric realism by hand-crafted transformations that involve complicated parameters and are prone to various unnatural alignments. The appearance realism is handled by different blending techniques where features are manually selected and still susceptible to artifacts. Our proposed technique instead adopts a GAN structure that learn geometry and appearance features from real images with little supervision, minimizing various inconsistency and artifacts effectively.

2.2. GAN

GANs [9] have achieved great success in generating realistic new images from either existing images or random noises. The main idea is to have a continuing adversarial learning between a generator and a discriminator, where the generator tries to generate more realistic images while the discriminator aims to distinguish the newly generated images from real images. Starting from generating MNIST handwritten digits, the quality of GAN-synthesized images has been improved greatly by the laplacian pyramid of adversarial networks [6]. This is followed by various efforts that employ a DNN architecture [33], stacking a pair of generators [52], learning more interpretable latent representations [4], adopting an alternative training method [1], etc.

Most existing GANs work towards synthesis realism in the appearance space. For example, CycleGAN [55] uses cycle-consistent adversarial networks for realistic image-

to-image translation, and so other relevant GANs [16, 37]. LR-GAN [48] generates new images by applying additional spatial transformation networks (STNs) to factorize shape variations. GP-GAN [46] composes high-resolution images by using Poisson blending [31]. A few GANs have been reported in recent years for geometric realsim, e.g., [21] presents a spatial transformer GAN (ST-GAN) by embedding STNs in the generator for geometric realism, [2] designs Compositional GAN that employs a self-consistent composition-decomposition network.

Most existing GANs synthesize images in either geometry space (e.g. ST-GAN) or appearance space (e.g. CycleGAN) but few in both spaces. In addition, the GAN-synthesized images are usually not suitable for training deep network models due to the lack of annotation or synthesis realism. Our proposed SF-GAN can achieve both appearance and geometry realism by synthesizing images in appearance and geometry spaces concurrently. Its synthesized images can be directly used to train more powerful deep network models due to their high realism.

2.3. Guided Filter

Guided Filters [12, 13] use one image as guidance for filtering another image which has shown superior performance in detail-preserving filtering. The filtering output is a linear transform of the guidance image by considering its structures, where the guidance image can be the input image itself or another different image. Guided filtering has been used in various computer vision tasks, e.g., [20] uses it for weighted averaging and image fusion, [53] uses a rolling guidance for fully-controlled detail smoothing in an iterative manner, [45] uses a fast guided filter for efficient image super-solution, [24] uses guided filters for high-quality depth map restoration, [23] uses guided filtering for tolerance to heavy noises and structure inconsistency, and [11] puts Guided filtering as a nonconvex optimization problem and proposes solutions via majorize-minimization [15].

Most GANs for image-to-image-translation can synthesize high-resolution images but the appearance transfer often suppresses image details such as edges and texture. How to keep the details of the original image while learning the appearance of the target remain an active research area. The proposed SF-GANs introduces guided filters into a cycle network which is capable of achieving appearance transfer and detail preserving concurrently.

3. The Proposed Method

The proposed SF-GAN consists of a geometry synthesizer and an appearance synthesizer, and the whole network is end-to-end trainable as illustrated in Fig. 2. Detailed network structure and training strategy will be introduced in the following subsections.

Table 1. The structure of the geometry estimation network within the STN in Fig. 2

Layers	Out Size	Configurations
Block1	16×50	3×3 $conv, 32, 2 \times 2$ $pool$
Block2	8×25	3×3 $conv, 64, 2 \times 2$ $pool$
Block3	4×13	3×3 $conv, 128, 2 \times 2$ $pool$
FC1	512	-
FC2	N	-

3.1. Geometry Synthesizer

The geometry synthesizer has a local GAN structure as highlighted by blue-color lines and boxes on the left of Fig. 2. It consists of a spatial transform network (STN), a composition module and a discriminator. The STN consists of an estimation network as shown in Table 1 and a transformation matrix which has N parameters that control the geometric transformation of the foreground object.

The foreground object and background image are concatenated to act as the input of the STN, where the estimation network will predict a transformation matrix to transform the foreground object. The transformation can be affine, homography, or thin plate spline [3] (We use thin plate spline for the scene text synthesis task and homography for the portrait wearing task). Each pixel in the transformed image is computed by applying a sampling kernel centered at a particular location in the original image. With pixels in the original and transformed images denoted by $P^s = (p_1^s, p_2^s, \ldots, p_N^s)$ and $P^t = (p_1^t, p_2^t, \ldots, p_N^t)$, we use a transformation matrix H to perform pixel-wise transformation as follows:

$$\begin{bmatrix} x_i^t \\ y_i^t \\ 1 \end{bmatrix} = H \begin{bmatrix} x_i^s \\ y_i^s \\ 1 \end{bmatrix} \tag{1}$$

where $p_i^s = (x_i^s, y_i^s)$ and $p_j^t = (x_i^t, y_i^t)$ denote the coordinates of the i-th pixel within the original and transformed image, respectively.

The transformed foreground object can thus be placed into the background image to form an initially composed image (*Composed Image* in Fig. 2). The discriminator *D2* in Fig. 2 learns to distinguish whether the composed image is realistic with respect to a set of *Real Images*. On the other hand, our study shows that real images are not good references for training geometry synthesizer. The reason is real images are realistic in both geometry and appearance spaces while the geometry can only achieve realism in geometry space. The difference in appearance space between the synthesized images and real images will mislead the training of geometry synthesizer. For optimal training of geometry synthesizer, the reference images should be realistic in the geometry space only and concurrently have similar appearance (e.g. colors and styles) with the initially composed

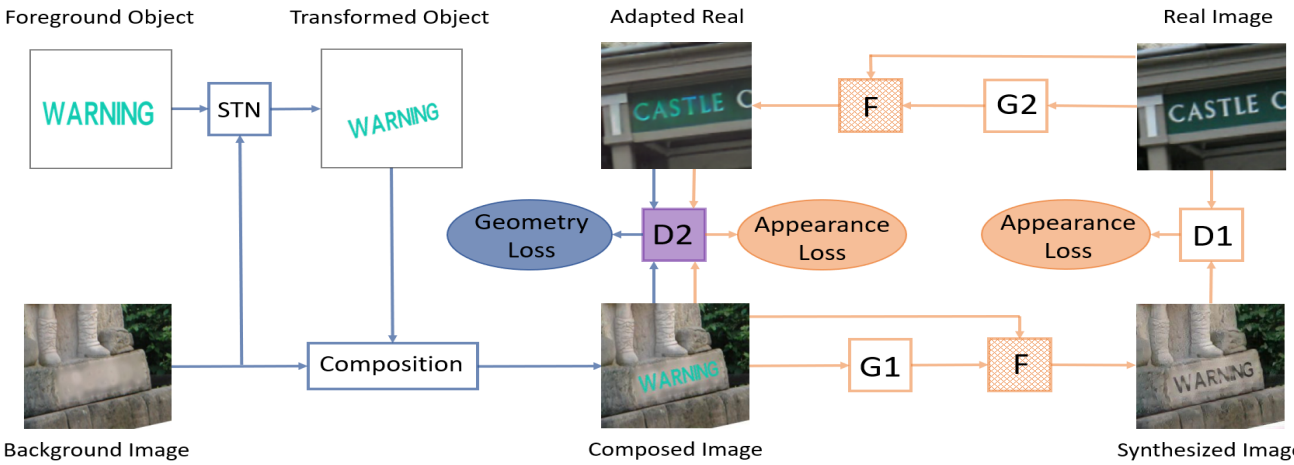

Figure 2. The structure of the proposed SF-GAN: The geometry synthesizer is highlighted by blue-color lines and boxes on the left and the appearance synthesizer is highlighted by orange-color lines and boxes on the right. *STN* denotes spatial transformation network, *F* denotes guided filters, *G1*, *G2*, *D1* and *D2* denote the generators and discriminators. For clarity, cycle loss and identity loss are not included.

images. Such reference images are difficult to create manually. In the SF-GAN, we elegantly use images from the appearance synthesizer (*Adapted Real* shown in Fig. 2) as the reference to train the geometry synthesizer, more details about the appearance synthesizer to be discussed in the following subsection.

3.2. Appearance Synthesizer

The appearance synthesizer is designed in a cycle structure as highlighted in orange-color lines and boxes on the right of Fig. 2. It aims to fuse the foreground object and background image to achieve synthesis realism in the appearance space. Image-to-image translation GANs also strive for realistic appearance but they usually lose visual details while performing the appearance transfer. Within the proposed SF-GAN, guided filters are introduced which help to preserve visual details effectively while working towards synthesis realism within the appearance space.

3.2.1 Cycle Structure

The proposed SF-GAN adopts a cycle structure for mapping between two domains, namely, the composed image domain and the real image domain. Two generators *G1* and *G2* are designed to achieve image-to-image translation in two reverse directions, *G1* from *Composed Image* to *Final Synthesis* and *G2* from *Real Images* to *Adapted Real* as illustrated in Fig. 2. Two discriminator *D1* and *D2* are designed to discriminate real images and translated images.

In particular, *D1* will strive to distinguish the adapted composed images (i.e. the *Composed Image* after domain adaptation by *G1*) and *Real Images*, forcing *G1* to learn to map from the *Composed Image* to *Final Synthesis* images that are realistic in the appearance space G2 will learn to map from *Real Images* to *Adapted Real* images, the images that ideally are realistic in the geometry space only but have similar appearance as the *Composed Image*. As discussed in the previous subsection, the *Adapted Real* from *G2* will be used as references for training the geometry synthesizer as it will better focus on synthesizing images with realistic geometry (as the interfering appearance difference has been compressed in the *Adapted Real*).

Image appearance transfer usually comes with detail loss. We address this issue from two perspectives. The fist is by adaptive combination of cycle loss and identity loss. Specifically, we adopt a weighted combination strategy that assigns higher weight to the cycle-loss for interested image regions while higher weight to the identify-loss for non-interested regions. Take scene text image synthesis as an example. By assigning a larger cycle-loss weight and smaller identity-loss to text regions, it ensures a multi-mode mapping of the text style while keeping the background similar to the original image. The second is by introducing guided filters into the cycle structure for detail preserving, more details to be described in the next subsection.

3.2.2 Guided Filter

Guided filter was designed to perform edge-preserving image smoothing. It influences the filtering by using structures in a guidance image As appearance transfer in most image-to-image-translation GANs tends to lose image details, we introduce guided filters (*F* as shown in Fig. 2) into the SF-GAN for detail preserving within the translated images. The target is to perform appearance transfer on the foreground object (within the *Composed Image*) only while keeping the background image with minimum changes.

We introduce guided filters into the proposed SF-GAN and formulate the detail-preserving appearance transfer as a joint up-sampling problem as illustrated in Fig. 3. In par-

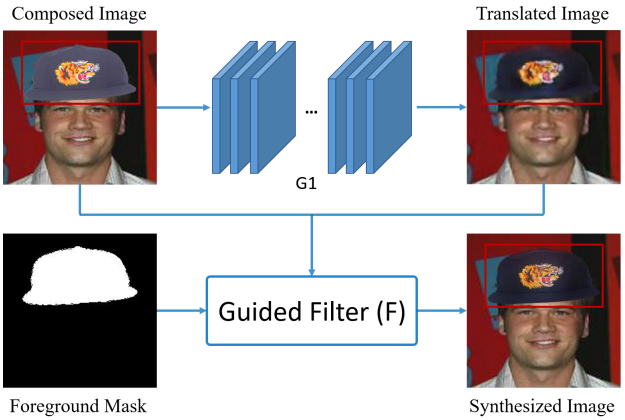

Figure 3. Detailed structure of the guided filter *F*: Given an image to be filtered (*Composed Image* in Fig. 2), a translated image with smoothed details (the output of *G1* in Fig. 2 where details are lost around background face and foreground hat areas) and the mask of foreground object hat (provided), *F* produces a new image with full details (*Synthesized Image*, the output of *F* at the bottom in Fig. 2). It can be seen that the guided filter preserves details of both background image (e.g. the face area) and foreground hat (e.g. the image areas highlighted by the red-color box).

ticular, the translated images from the output of *G1* (image details lost) is the input image I to be filtered and the initially *Composed Image* (image details unchanged) shown in Fig. 2 acts as the guidance image R to provide edge and texture details. The detail-preserving image T (corresponding to the *Synthesized Image* in Fig. 2) can thus be derived by minimizing the reconstruction error between I and T, subjects to a linear model:

$$T_i = a_k I_i + b_k, \forall i \in \omega_k \tag{2}$$

where i is the index of a pixel and ω_k is a local square window centered at pixel k.

To determine the coefficients of the linear model a_k and b_k, we seek a solution that minimizes the difference between T and the filter input R which can be derived by minimizing the following cost function in the local window:

$$E(a_k, b_k) = \sum_{i \in \omega_k} ((a_k I_i + b_k - R_i)^2 + \epsilon a_k^2) \tag{3}$$

where ϵ is a regularization parameter that prevents a_k from being too large. It can be solved via linear regression:

$$a_k = \frac{\frac{1}{|\omega|}\sum_{i \in \omega_k} I_i - \mu_k \overline{R}_k}{\overline{\sigma}_k + \epsilon} \tag{4}$$

$$b_k = \overline{R}_k - a_k \mu_k \tag{5}$$

where μ_k and σ_k^2 are the mean and variance of I in ω_k, $|\omega|$ is the number of pixels in ω_k, and $\overline{R}_k = \frac{1}{|\omega|}\sum_{i \in \omega_k}$ is the mean of R in ω_k.

By applying the linear model to all windows ω_k in the image and computing (a_k, b_k), the filter output can be derived by averaging all possible values of T_i:

$$T_i = \frac{1}{|\omega|} \sum_{k:i \in \mu_k} (a_k I_i + bk) = \overline{a}_i I_i + \overline{b}_i \tag{6}$$

where $\overline{a}_i = \frac{1}{|\omega|}\sum_{k \in \omega_k} a_k$ and $\overline{b}_i = \frac{1}{|\omega|}\sum_{k \in \omega_i} b_k$. We integrate the guide filter into the cycle structure network to implement an end-to-end trainable system.

3.3. Adversarial Training

The proposed SF-GAN is designed to achieve synthesis realism in both geometry and appearance spaces. The SF-GAN training therefore has two adversarial objectives, one is to learn the real geometry and the other is to learn the real appearance The geometry synthesizer and appearance synthesizer are actually two local GANs that are interconnected and need coordination during the training. For presentation clarity, we denote the *Foreground Object* and *Background Image* in Fig. 2 as the x, the *Composed Image* as y and the *Real Image* as z which belongs to domains X, Y and Z, respectively.

For the geometry synthesizer, the *STN* can actually be viewed as a generator $G0$ which predicts transportation parameters for x. After the transformation of the *Foreground Object* and *Composition*, the *Composed Image* becomes the input of the discriminator $D2$ and the training reference $z^{'}$ comes from $G2(z)$ of the appearance synthesizer. For the geometry synthesizer, we adopt the Wasserstein GAN [ ] objective for training which can be denoted by:

$$\min_{G0} \max_{D2} E_{x \sim X}[D2(G0(x))] - E_{z' \sim Z'}[D2(z)] \tag{7}$$

where $Z^{'}$ denotes the domains for $z^{'}$. Since *G0* aims to minimize this objective against an adversary *D2* that tries to maximize it, the loss functions of *D2* and *G0* can be defined by:

$$L_{D2} = E_{x \sim X}[D2(G0(x)] - E_{z' \sim Z'}[D2(z^{'})] \tag{8}$$

$$L_{G0} = -E_{x \sim X}[D2(G0(x))] \tag{9}$$

The appearance synthesizer adopts a cycle structure that consists of two mappings $G1 : Y \rightarrow Z$ and $G2 : Z \rightarrow Y$. It has two adversarial discriminators *D1* and *D2*. D2 is shared between the geometry and appearance synthesizers, and it aims to distinguish y from *G2(z)* within the appearance synthesizer. The learning objectives thus consists of an adversarial loss for the mapping between domains and a cycle consistency loss for preventing the mode collapse. For the adversarial loss, the objective of the mapping $G1 : Y \rightarrow Z$ (and the same for the reverse mapping $G2 : Z \rightarrow Y$) can be defined by:

$$L_{D1} = E_{y \sim Y}[D1(G1(y)] - E_{z \sim Z}[D2(z)] \tag{10}$$

$$L_{G1} = -E_{y\sim Y}[D1(G1(y))] \quad (11)$$

As the adversarial losses cannot guarantee that the learned function maps an individual input y to a desired output z, we introduce cycle-consistency, aiming to ensure that the image translation cycle will bring x back to the original image, i.e. $y \rightarrow G1(y) \rightarrow G2(G1(y)) = y$. The cycle-consistency can be achieved by a cycle-consistency loss:

$$L_{G1_{cyc}} = E_{y\sim p(y)}[\|G2(G1(y)) - y\|] \quad (12)$$

$$L_{G2_{cyc}} = E_{z\sim p(z)}[\|G1(G2(z)) - z\|] \quad (13)$$

We also introduce the identity loss to ensure that the translated image preserves features of the original image:

$$L_{G1_{idt}} = E_{y\sim Y}[\|G1(y) - y\|] \quad (14)$$

$$L_{G2_{idt}} = E_{z\sim Z}[\|G2(z) - z\|] \quad (15)$$

For each training step, the model needs to update the geometry synthesizer and appearance synthesizer separately. In particular, L_{D2} and L_{G0} are optimized alternately while updating the geometry synthesizer. While updating the appearance synthesizer, all weights of the geometry synthesizer are freezed. In the mapping $G1 : Y \rightarrow Z$, L_{D1} and $L_{G1} + \lambda_1 L_{G1_{cyc}} + \lambda_2 L_{G1_{idt}}$ are optimized alternately where λ_1 and λ_2 controls the relative importance of the cycle-consistency loss and the identity loss, respectively. In the mapping $G2 : Z \rightarrow Y$, L_{D2} and $L_{G2} + \lambda_1 L_{G2_{cyc}} + \lambda_2 L_{G2_{idt}}$ are optimized alternately.

It should be noted that the sequential updating is necessary for end-to-end training of the proposed SF-GAN. If discarding the geometry loss, we need update the geometry synthesizer according to the loss function of the appearance synthesizer. On the other hand, the appearance synthesizer will generate blurry foreground objects regardless of the geometry synthesizer and this is similar to GANs for direct image generation. As discussed before, the direct image generation cannot provide accurate annotation information and the directly generated images also have low quality and are not suitable for training deep network models.

4. Experiments

4.1. Datasets

ICDAR2013 [19] is used in the Robust Reading Competition in the International Conference on Document Analysis and Recognition (ICDAR) 2013. It contains 848 word images for network training and 1095 for testing.

ICDAR2015 [18] is used in the Robust Reading Competition under ICDAR 2015. It contains incidental scene text images that are captured without preparation before capturing. 2077 text image patches are cropped from this dataset, where a large amount of cropped scene texts suffer from perspective and curvature distortions.

IIIT5K [28] has 2000 training images and 3000 test images that are cropped from scene texts and born-digital images. Each word in this dataset has a 50-word lexicon and a 1000-word lexicon, where each lexicon consists of a ground-truth word and a set of randomly picked words.

SVT [44] is collected from the Google Street View images that were used for scene text detection research. 647 words images are cropped from 249 street view images and most cropped texts are almost horizontal.

SVTP [32] has 639 word images that are cropped from the SVT images. Most images in this dataset suffer from perspective distortion which are purposely selected for evaluation of scene text recognition under perspective views.

CUTE [35] has 288 word images mose of which are curved. All words are cropped from the CUTE dataset which contains 80 scene text images that are originally collected for scene text detection research.

CelebA [25] is a face image dataset that consists of more than 200k celebrity images with 40 attribute annotations. This dataset is characterized by large quantities, large face pose variations, complicated background clutters, rich annotations, and it is widely used for face attribute prediction.

4.2. Scene Text Synthesis

Data Preparation: The SF-GAN needs a set of *Real Images* to act at references as illustrated in Fig. 2. We create the *Real Images* by cropping the text image patches from the training images of ICDAR2013 [19], ICDAR2015 [18] and SVT [44] by using the provided annotation boxes. While cropping the text image patches, we extend the annotation box (by an extra 1/4 of the width and height of the annotation boxes) to include certain local geometric structures

Besides the *Real Images*, SF-GAN also needs a set of *Background Images* as shown in Fig. 2. For scene text image synthesis, we collect the background images by smoothing out the text pixels of the cropped *Real Images*. Further, the *Foreground Object* (text for scene text synthesis) is computer-generated by using a 90k-lexicon. The created *Background Images*, *Foreground Texts* and *Real Images* are fed to the network to train the SF-GAN.

For the training of scene text recognition model, texts need to be cropped out with tighter boxes (to exclude extra background). With the text maps as denoted by *Transformed Object* in Fig. 2, scene text patches can be cropped out accurately by detecting a minimal external rectangle.

Results Analysis: We use 1 million SF-GAN synthesized scene text images to train scene text recognition models and use the model recognition performance to evaluate the usefulness of the synthesized images. In addition, the SF-GAN is benchmarked with a number of state-of-the-art synthe-

Table 2. Scene text recognition accuracy over the datasets ICDAR2013, ICDAR2015, SVT, IIIT5K, SVTP and CUTE, where 1 million synthesized text images are used for all comparison methods as as listed.

Methods	ICDAR2013	ICDAR2015	SVT	IIIT5K	SVTP	CUTE	AVERAGE
Jaderberg [17]	58.1	35.5	67.0	57.2	**48.9**	35.3	50.3
Gupta [10]	62.2	38.2	48.8	59.1	38.9	36.3	47.3
Zhan [50]	**62.5**	37.7	63.5	59.5	46.7	36.9	51.1
ST-GAN [21]	57.2	35.3	63.8	57.3	43.2	34.1	48.5
SF-GAN(BS)	55.9	34.9	64.0	55.4	42.8	33.7	47.8
SF-GAN(GS)	57.3	35.6	66.5	57.7	43.9	36.1	49.5
SF-GAN(AS)	58.1	36.4	66.7	58.5	45.3	35.7	50.1
SF-GAN	61.8	**39.0**	**69.3**	**63.0**	48.6	**40.6**	**53.7**

Figure 4. Illustration of scene text image synthesis by different GANs: Rows 1-2 are foreground texts and background images as labelled. Rows 3-4 show the images synthesized by ST-GAN and CycleGAN, respectively. Row 5 shows images synthesized by SF-GAN(GS), the output of the geometry synthesizer in SF-GAN (*Composed Image* in Fig. 2). The last row shows images synthesized by the proposed SF-GAN.

sis techniques by randomly selecting 1 million synthesized scene text images from [17] and randomly cropping 1 million scene text images from [10] and [50]. Beyond that, we also synthesize 1 million scene text images with random text appearance by using ST-GAN [21]. There are many scene text recognition models [38, 39, 41, 36, 5], we design an attentional scene text recognizer with a 50-layer ResNet as the backbone network.

For ablation analysis, we evaluate SF-GAN(GS) which denotes the output of the geometry synthesizer (*Composed Image* as shown in Fig. 2) and SF-GAN(AS) which denotes the output of the appearance synthesizer with random geometric alignments. A baseline SF-GAN (BS) is also trained where texts are placed with random alignment and appearance. The three SF-GANs also synthesize 1 million images each for scene text recognition tests. The recognition tests are performed over four regular scene text datasets ICDAR2013 [19], ICDAR2015 [18], SVT [44], IIIT5K [28] and two irregular datasets SVTP [32] and CUTE [35] as described in **Datasets**. Besides the scene text recognition, we also perform user studies with Amazon Mechanical Turk (AMT) where users are recruited to tell whether SF-GAN synthesized images are real or synthesized.

Tables 2 and 3 show scene text recognition and AMT user study results. As Table 2 shows, SF-GAN achieves the highest recognition accuracy for most of the 6 datasets and an up to 3% improvement in average recognition accuracy (across the 6 datasets), demonstrating the superior usefulness of its synthesized images while used for training scene text recognition models. The ablation study shows that the proposed geometry synthesizer and appearance synthesizer both help to synthesize more realistic and useful image in recognition model training. In addition, they are complementary and their combination achieves a 6% improvement in average recognition accuracy beyond the baseline SF-GAN(BS). The AMT results in the second column of Table 3 also show that the SF-GAN synthesized scene text images are much more realistic than state-of-the-art synthesis techniques. Note the synthesized images by [17] are gray-scale and not included in the AMT user study.

Fig. 4 shows a few synthesis images by using the proposed SF-GAN and a few state-of-the-art GANs. As Fig. 4 shows, ST-GAN can achieve geometric alignment but the appearance is clearly unrealistic within the synthesized im-

Table 3. AMT user study to evaluate the realism of synthesized images. Percentages represent the how often the images in each category were classified as real by Turkers.

Methods	Text	Glass	Hat
Gupta [10]	38.0	-	-
Zhan [50]	41.5	-	-
ST-GAN [21]	31.6	41.7	42.6
Real	74.1	78.6	78.2
SF-GAN	57.7	62.0	67.3

ages. The CycleGAN can adapt the appearance of the foreground texts to certain degrees but it ignores real geometry. This leads to not only unrealistic geometry but also degraded appearance as the discriminator can easily distinguish generated images and real images according to the geometry difference. The SF-GAN (GS) gives the output of the geometry synthesizer, i.e. the *Composed Image* as shown in Fig. 2, which produces better alignment due to good references from the appearance synthesizer. In addition, it can synthesize curve texts due to the use of a thin plate spline transformation [3]. The fully implemented SF-GAN can further learn text appearance from real images and synthesize highly realistic scene text images. Besides, we can see that the proposed SF-GAN can learn from neighboring texts within the background images and adapt the appearance of the foreground texts accordingly.

4.3. Portrait Wearing

Data preparation: We use the dataset CelebA [25] and follow the provided training/test split for portrait wearing experiment. The training set is divided into two groups by using the annotation 'glass' and 'hat', respectively. For the glass case, one group of people with glasses serve as the real data for matching against in our adversarial settings and the other group without glasses serves as the background. For the foreground glasses, we crop 15 pairs of front-parallel glasses and reuse them to randomly compose with the background images. According to our experiment, 15 pairs of glasses as the foreground objects are sufficient to train a robust model. The hat case has the similar setting, except that we use 30 cropped hats as the foreground objects.

Results Analysis: Fig 5. shows a few SF-GAN synthesized images and comparisons with ST-GAN synthesized images. As Fig. 5 shows, ST-GAN achieves realism in the geometry space by aligning the glasses and hats with the background face images. On the other hand, the synthesized images are unrealistic in the appearance space with clear artifacts in color, contrast and brightness. As a comparison, the SF-GAN synthesized images are much more realistic in both geometry and appearance spaces. In particular, the foreground glasses and hats within the SF-GAN

Figure 5. Illustration of portrait-wearing by different GANs: Columns 1-2 show foreground hats and glasses and background face images, respectively. Columns 3-4 show images synthesized by by ST-GAN [21] and our proposed SF-GAN, respectively.

synthesized images have harmonious brightness, contrast, and blending with the background face images. Additionally, the proposed SF-GAN also achieve better geometric alignment as compared with ST-GAN which focuses on geometric alignment only. We conjecture that the better geometric alignment is largely due to the reference from the appearance synthesizer. The AMT results as shown in the last two columns of Table 3 also show the superior synthesis performance of our proposed SF-GAN.

5. Conclusions

This paper presents a SF-GAN, an end-to-end trainable network that synthesize realistic images given foreground objects and background images. The SF-GAN is capable of achieving synthesis realism in both geometry and appearance spaces concurrently. The first scene text image synthesis study shows that the proposed SF-GAN is capable of synthesizing useful images to train better recognition models. The second portrait-wearing study shows the SF-GAN is widely applicable and can be easily extend to other tasks. We will continue to study SF-GAN for full-image synthesis for training better detection models.

References

[1] Martin Arjovsky, Soumith Chintala, and Lon Bottou. Wasserstein generative adversarial networks. In *ICML*, 2017. 1, 2, 5

[2] Samaneh Azadi, Deepak Pathak, Sayna Ebrahimi, and Trevor Darrell. Compositional gan: Learning conditional image composition. *arXiv:1807.07560*, 2018. 1, 3

[3] Fred L. Bookstein. Principal warps: Thin-plate splines and the decomposition of deformations. *TPAMI*, 11(6), 1989. 3, 8

[4] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *NIPS*, 2016. 2

[5] Zhanzhan Cheng, Yangliu Xu, Fan Bai, Yi Niu, Shiliang Pu, and Shuigeng Zhou. Aon: Towards arbitrarily-oriented text recognition. In *CVPR*, 2018. 7

[6] Emily Denton, Soumith Chintala, Arthur Szlam, and Rob Fergus. Deep generative image models using a laplacian pyramid of adversarial networks. In *NIPS*, 2015. 2

[7] Debidatta Dwibedi, Ishan Misra, and Martial Hebert. Cut, paste and learn: Surprisingly easy synthesis for instance detection. In *ICCV*, 2017. 2

[8] Adrien Gaidon, Qiao Wang, Yohann Cabon, and Eleonora Vig. Virtual worlds as proxy for multi-object tracking analysis. In *CVPR*, 2016. 2

[9] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. In *NIPS*, pages 2672–2680, 2014. 1, 2

[10] Ankush Gupta, Andrea Vedaldi, and Andrew Zisserman. Synthetic data for text localisation in natural images. In *CVPR*, 2016. 2, 7, 8

[11] Bumsub Ham, Minsu Cho, and Jean Ponce. Robust guided image filtering using nonconvex potentials. *TPAMI*, 2018. 3

[12] Kaiming He and Jian Sun. Fast guided filter. *arXiv:1505.00996*, 2015. 3

[13] Kaiming He, Jian Sun, and Xiaoou Tang. Guided image filtering. *TPAMI*, 2013. 3

[14] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Alexei A. Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *ICML*, 2018. 1, 2

[15] David R. Hunter and Kenneth Lange. Quantile regression via an mm algorithm. *Journal of Computational and Graphical Statistics*, 2000. 3

[16] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017. 1, 2, 3

[17] Max Jaderberg, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Synthetic data and artificial neural networks for natural scene text recognition. In *NIPS Deep Learning Workshop*, 2014. 2, 7

[18] Dimosthenis Karatzas, Lluis Gomez-Bigorda, Anguelos Nicolaou, Suman Ghosh, Andrew Bagdanov, Masakazu Iwamura, Jiri Matas, Lukas Neumann, Vijay Ramaseshan Chandrasekhar, Shijian Lu, Faisal Shafait, Seiichi Uchida, and Ernest Valveny. Icdar 2015 competition on robust reading. In *ICDAR*, pages 1156–1160, 2015. 6, 7

[19] D. Karatzas, F. Shafait, S. Uchida, M. Iwamura, S. R. Mestre, J. Mas, D. F. Mota, J. A. Almazan, L. P. de las Heras, and et al. Icdar 2013 robust reading competition. In *ICDAR*, pages 1484–1493, 2013. 6, 7

[20] Shutao Li, Xudong Kang, and Jianwen Hu. Image fusion with guided filtering. *TIP*, 22(7), 2013. 3

[21] Chen-Hsuan Lin, Ersin Yumer, Oliver Wang, Eli Shechtman, and Simon Lucey. St-gan: Spatial transformer generative adversarial networks for image compositing. In *CVPR*, 2018. 1, 3, 7, 8

[22] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *NIPS*, 2017. 1, 2

[23] Wei Liu, Xiaogang Chen, Chunhua Shen, Jingyi Yu, Qiang Wu, and Jie Yang. Robust guided image filtering. *arXiv:1703.09379*, 2017. 3

[24] Wei Liu, Yun Gu, Chunhua Shen, Xiaogang Chen, Qiang Wu, and Jie Yang. Data driven robust image guided depth map restoration. *TIP*, 2017. 3

[25] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, 2015. 6, 8

[26] Fujun Luan, Sylvain Paris, Eli Shechtman, and Kavita Bala. Deep painterly harmonization. *arXiv:1804.03189*, 2018. 2

[27] Simon Osindero Mehdi Mirza. Conditional generative adversarial nets. *arXiv:1411.1784*, 2014. 1

[28] Anand Mishra, Karteek Alahari, and C.V. Jawahar. Scene text recognition using higher order language priors. In *BMVC*, 2012. 6, 7

[29] Yair Movshovitz-Attias, Takeo Kanade, and Yaser Sheikh. How useful is photo-realistic rendering for visual learning? In *ECCV*, 2016. 2

[30] Dennis Park and Deva Ramanan. Articulated pose estimation with tiny synthetic videos. In *CVPR*, 2015. 2

[31] Patrick Pérez, Michel Gangnet, and Andrew Blake. Poisson image editing. *TOG*, 22(3), 2003. 2, 3

[32] Trung Quy Phan, Palaiahnakote Shivakumara, Shangxuan Tian, and Chew Lim Tan. Recognizing text with perspective distortion in natural scenes. In *ICCV*, 2013. 6, 7

[33] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *ICLR*, 2016. 1, 2

[34] Stephan R. Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *ECCV*, 2016. 2

[35] Anhar Risnumawan, Palaiahankote Shivakumara, Chee Seng Chan, and Chew Lim Tan. A robust arbitrary text detection system for natural scene images. *Expert Syst. Appl.*, 41(18):8027–8048, 2014. 6, 7

[36] Baoguang Shi, Mingkun Yang, Xinggang Wang, Pengyuan Lyu, Cong Yao, and Xiang Bai. Aster: An attentional scene text recognizer with flexible rectification. *TPAMI*, 2018. 7

[37] Ashish Shrivastava, Tomas Pfister, Oncel Tuzel, Josh Susskind, Wenda Wang, and Russ Webb. Learning from simulated and unsupervised images through adversarial training. In *CVPR*, 2017. 3

[38] Bolan Su and Shijian Lu. Accurate scene text recognition based on recurrent neural network. In *ACCV*, 2014. 7

[39] Bolan Su and Shijian Lu. Accurate recognition of words in scenes without character segmentation using recurrent neural network. *PR*, pages 397–405, 2017. 7

[40] Hao Su, Charles R. Qi, Yangyan Li, and Leonidas Guibas. Render for cnn: Viewpoint estimation in images using cnns trained with rendered 3d model views. In *ICCV*, 2015. 2

[41] Shangxuan Tian, Ujjwal Bhattacharya, Shijian Lu, Bolan Su, Qingqing Wang, Xiaohua Wei, Yue Lu, and Chew Lim Tan. Multilingual scene character recognition with cooccurrence of histogram of oriented gradients. *PR*, pages 125–134, 2016. 7

[42] Yi-Hsuan Tsai, Xiaohui Shen, Zhe Lin, Kalyan Sunkavalli, Xin Lu, and Ming-Hsuan Yang. Deep image harmonization. In *CVPR*, 2017. 2

[43] Matthew Uyttendaele, Ashley Eden, and Richard Szeliski. Eliminating ghosting and exposure artifacts in image mosaics. In *CVPR*, 2001. 2

[44] K. Wang, B. Babenko, and S. Belongie. End-to-end scene text recognition. In *ICCV*, 2011. 6, 7

[45] Huikai Wu, Shuai Zheng, Junge Zhang, and Kaiqi Huang. Fast end-to-end trainable guided filter. In *CVPR*, 2017. 3

[46] Huikai Wu, Shuai Zheng, Junge Zhang, and Kaiqi Huang. Gp-gan: Towards realistic high-resolution image blending. *arXiv:1703.07195*, 2017. 3

[47] Chuhui Xue, Shijian Lu, and Fangneng Zhan. Accurate scene text detection through border semantics awareness and bootstrapping. In *ECCV*, pages 355–372, 2018. 2

[48] Jianwei Yang, Anitha Kannan, Dhruv Batra, and Devi Parikh. Lr-gan: Layered recursive generative adversarial networks for image generation. In *ICLR*, 2017. 3

[49] Fangneng Zhan and Shijian Lu. Esir: End-to-end scene text recognition via iterative image rectification. In *CVPR*, 2019. 2

[50] Fangneng Zhan, Shijian Lu, and Chuhui Xue. Verisimilar image synthesis for accurate detection and recognition of texts in scenes. In *ECCV*, pages 249–266, 2018. 2, 7, 8

[51] Fangneng Zhan, Hongyuan Zhu, and Shijian Lu. Scene text synthesis for efficient and effective deep network training. *arXiv:1901.09193*, 2019. 2

[52] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *ICCV*, 2017. 2

[53] Qi Zhang, Xiaoyong Shen, Li Xu, and Jiaya Jia. Rolling guidance filter. In *ECCV*, 2014. 3

[54] Yinan Zhao, Brian Price, Scott Cohen, and Danna Gurari. Guided image inpainting: Replacing an image region by pulling content from another image. *arXiv:1803.08435*, 2018. 2

[55] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017. 1, 2