Unsupervised Holistic Image Generation from Key Local Patches

Donghoon Lee[1], Sangdoo Yun[2], Sungjoon Choi[1], Hwiyeon Yoo[1], Ming-Hsuan Yang[3,4], and Songhwai Oh[1]

[1] Electrical and Computer Engineering and ASRI, Seoul National University
[2] Clova AI Research, NAVER
[3] Electrical Engineering and Computer Science, University of California at Merced
[4] Google Cloud AI

Abstract. We introduce a new problem of generating an image based on a small number of key local patches without any geometric prior. In this work, key local patches are defined as informative regions of the target object or scene. This is a challenging problem since it requires generating realistic images and predicting locations of parts at the same time. We construct adversarial networks to tackle this problem. A generator network generates a fake image as well as a mask based on the encoder-decoder framework. On the other hand, a discriminator network aims to detect fake images. The network is trained with three losses to consider spatial, appearance, and adversarial information. The spatial loss determines whether the locations of predicted parts are correct. Input patches are restored in the output image without much modification due to the appearance loss. The adversarial loss ensures output images are realistic. The proposed network is trained without supervisory signals since no labels of key parts are required. Experimental results on seven datasets demonstrate that the proposed algorithm performs favorably on challenging objects and scenes.

Keywords: Image synthesis · Generative adversarial networks

1 Introduction

The goal of image generation is to construct images that are as barely distinguishable from target images which may contain general objects, diverse scenes, or human drawings. Synthesized images can contribute to a number of applications such as the image to image translation [7], image super-resolution [13], 3D object modeling [36], unsupervised domain adaptation [15], domain transfer [39], future frame prediction [33], image inpainting [38], image editing [43], and feature recovering of astrophysical images [29].

In this paper, we introduce a new image generation problem: a holistic image generation conditioned on a small number of local patches of objects or scenes without any geometry prior. It aims to estimate what and where object parts are needed to appear and how to fill in the remaining regions. There are various

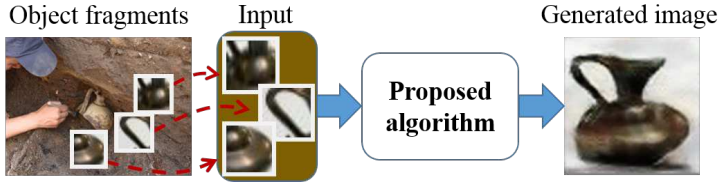

Fig. 1. The proposed algorithm is able to synthesize an image from key local patches without geometric priors, e.g., restoring broken pieces of ancient ceramics found in ruins. Convolutional neural networks are trained to predict locations of input patches and generate the entire image based on adversarial learning.

applications for this problem. For example, in a surveillance system, objects are often occluded and we need to recover the whole appearance from limited information. For augmented reality, by rendering plausible scenes based on a few objects, the experience of users become more realistic and diverse. Combining parts of different objects can generate various images in a target category, e.g., designing a new car based on parts of BMW and Porsche models. Broken objects that have missing parts can be restored as shown in Figure 1. While the problem is related to image completion and scene understanding tasks, it is more general and challenging than each of these problems due to following reasons.

First, spatial arrangements of input patches need to be inferred since the data does not contain explicit information about the location. To tackle this issue, we assume that inputs are key local patches which are informative regions of the target image. Therefore, the algorithm should learn the spatial relationship between key parts of an object or scene. Our approach obtains key regions without any supervision such that the whole algorithm is developed within the unsupervised learning framework.

Second, we aim to generate an image while preserving the key local patches. As shown in Figure 1, the appearances of input patches are included in the generated image without significant modification. In other words, the inputs are not directly copied to the output image. It allows us to create images more flexibly such that we can combine key patches of different objects as inputs. In such cases, input patches must be deformed by considering each other.

Third, the generated image should look closely to a real image in the target category. Unlike the image inpainting problem, which mainly replaces small regions or eliminates minor defects, our goal is to reconstruct a holistic image based on limited appearance information contained in a few patches.

To address the above issues, we adopt the adversarial learning scheme [4] in this work. The generative adversarial network (GAN) contains two networks which are trained based on the min-max game of two players. A generator network typically generates fake images and aims to fool a discriminator, while a discriminator network seeks to distinguish fake images from real images. In our case, the generator network is also responsible for predicting the locations of input patches. Based on the generated image and predicted mask, we design three

losses to train the network: a spatial loss, an appearance loss, and an adversarial loss, corresponding to the aforementioned issues, respectively.

While a conventional GAN is trained in an unsupervised manner, some recent methods formulate it in a supervised manner by using labeled information. For example, a GAN is trained with a dataset that has 15 or more joint positions of birds [25]. Such labeling task is labor intensive since GAN-based algorithms need a large amount of training data to achieve high-quality results. In contrast, experiments on seven challenging datasets that contain different objects and scenes, such as faces, cars, flowers, ceramics, and waterfalls, demonstrate that the proposed unsupervised algorithm generates realistic images and predict part locations well. In addition, even if inputs contain parts from different objects, our algorithm is able to generate reasonable images.

The main contributions are as follows. First, we introduce a new problem of rendering realistic image conditioned on the appearance information of a few key patches. Second, we develop a generative network to jointly predict the mask and image without supervision to address the defined problem. Third, we propose a novel objective function using additional fake images to strengthen the discriminator network. Finally, we provide new datasets that contain challenging objects and scenes.

2 Related Work

Image Generation. Image generation is an important problem that has been studied extensively in computer vision. With the recent advances in deep convolutional neural networks [12,31], numerous image generation methods have achieved the state-of-the-art results. Dosovitskiy et al. [3] generate 3D objects by learning transposed convolutional neural networks. In [10], Kingma et al. propose a method based on variational inference for stochastic image generation. An attention model is developed by Gregor et al. [5] to generate an image using a recurrent neural network. Recently, the stochastic PixelCNN [21] and PixelRNN [22] are introduced to generate images sequentially.

The generative adversarial network [4] is proposed for generating sharp and realistic images based on two competing networks: a generator and a discriminator. Numerous methods [28,42] have been proposed to improve the stability of the GAN. Radford et al. [24] propose deep convolutional generative adversarial networks (DCGAN) with a set of constraints to generate realistic images effectively. Based on the DCGAN architecture, Wang et al. [34] develop a model to generate the style and structure of indoor scenes (SSGAN), and Liu et al. [15] present a coupled GAN which learns a joint distribution of multi-domain images, such as color and depth images.

Conditional GAN. Conditional GAN approaches [18,26,40] are developed to control the image generation process with label information. Mizra et al. [18] propose a class-conditional GAN which uses discrete class labels as the conditional information. The GAN-CLS [26] and StackGAN [40] embed a text describing an

image into the conditional GAN to generate an image corresponding to the condition. On the other hand, the GAWWN [25] creates numerous plausible images based on the location of key points or an object bounding box. In these methods, the conditional information, e.g., text, key points, and bounding boxes, is provided in the training data. However, it is labor intensive to label such information since deep generative models require a large amount of training data. In contrast, key patches used in the proposed algorithm are obtained without the necessity of human annotation.

Numerous image conditional models based on GANs have been introduced recently [13,43,39,38,23,14,30,7]. These methods learn a mapping from the source image to target domain, such as image super-resolution [13], user interactive image manipulation [43], product image generation from a given image [39], image inpainting [38,23], style transfer [14] and realistic image generation from synthetic image [30]. Isola et al. [7] tackle the image-to-image translation problem including various image conversion examples such as day image to night image, gray image to color image, and sketch image to real image, by utilizing the U-net [27] and GAN. In contrast, the problem addressed in this paper is the holistic image generation based on only a small number of local patches. This challenging problem cannot be addressed by existing image conditional methods as the domain of the source and target images are different.

Unsupervised Image Context Learning. Unsupervised learning of the spatial context in an image [2,20,23] has attracted attention to learn rich feature representations without human annotations. Doersch et al. [2] train convolutional neural networks to predict the relative position between two neighboring patches in an image. The neighboring patches are selected from a grid pattern based on the image context. To reduce the ambiguity of the grid, Noroozi et al. [20] divide the image into a large number of tiles, shuffle the tiles, and then learn a convolutional neural network to solve the jigsaw puzzle problem. Pathak et al. [23] address the image inpainting problem which predicts missing pixels in an image, by training a context encoder. Through the spatial context learning, the trained networks are successfully applied to various applications such as object detection, classification and semantic segmentation. However, discriminative models [2,20] can only infer the spatial arrangement of input patches, and the image inpainting method [23] requires the spatial information of the missing pixels. In contrast, we propose a generative model which is capable of not only inferring the spatial arrangement of inputs but also generating the entire image.

Image reconstruction from local information. Weinzaepfel et al. [35] reconstruct an image from local descriptors such as SIFT while the locations are known. This method retrieves an image patch for each region of interest from a database based on the similarity of local descriptors. These patches are then warped into a single image and stitched seamlessly. Zhang et al. [41] extrapolate an image from a limited field of view to a panoramic image. An input image is aligned with a guidance panorama image such that the unseen viewpoint is predicted based on self-similarity.

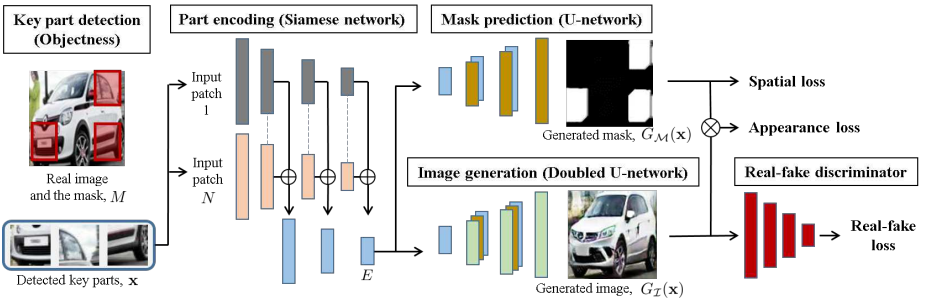

Fig. 2. Proposed network architecture. A bar represents a layer in the network. Layers of the same size and the same color have the same convolutional feature maps. Dashed lines in part encoding networks represent shared weights. An embedded vector is denoted as E.

3 Proposed Algorithm

Figure 2 shows the structure of the proposed network for image generation from a few patches. It is developed based on the concept of adversarial learning, where a generator and a discriminator compete with each other [4]. However, in the proposed network, the generator has two outputs: the predicted mask and generated image. Let $G_{\mathcal{M}}$ be a mapping from N observed image patches $\mathbf{x} = \{x_1, ..., x_N\}$ to a mask M, $G_{\mathcal{M}} : \mathbf{x} \rightarrow M$.[5] Also let $G_{\mathcal{I}}$ be a mapping from $\mathbf{x}$ to an output image y, $G_{\mathcal{I}} : \mathbf{x} \rightarrow y$. These mappings are performed based on three networks: a part encoding network, a mask prediction network, and an image generation network. The discriminator D is based on a convolutional neural network which aims to distinguish the real image from the image generated by $G_{\mathcal{I}}$. The function of each described module is essential in order to address the proposed problem. For example, it is not feasible to infer which region in the generated image should be similar to the input patches without the mask prediction network.

We use three losses to train the network. The first loss is the spatial loss $\mathcal{L}_S$. It compares the inferred mask and real mask which represents the cropped region of the input patches. The second loss is the appearance loss $\mathcal{L}_A$, which maintains input key patches in the generated image without much modification. The third loss is the adversarial loss $\mathcal{L}_R$ to distinguish fake and real images. The whole network is trained by the following min-max game:

$$\min_{G_{\mathcal{M}}, G_{\mathcal{I}}} \max_{D} \mathcal{L}_R(G_{\mathcal{I}}, D) + \lambda_1 \mathcal{L}_S(G_{\mathcal{M}}) + \lambda_2 \mathcal{L}_A(G_{\mathcal{M}}, G_{\mathcal{I}}), \tag{1}$$

where λ_1 and λ_2 are weights for the spatial loss and appearance loss, respectively.

[5] Here, $\mathbf{x}$ is a set of image patches resized to the same width and height suitable for the proposed network and N is the number of image patches in $\mathbf{x}$.

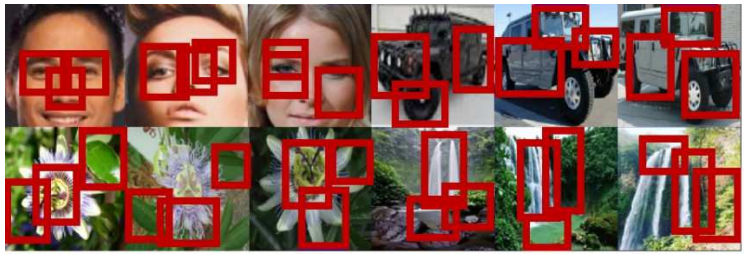

Fig. 3. Examples of detected key patches on faces [16], vehicles [11], flowers [19], and waterfall scenes. Three regions with top scores from the EdgeBox algorithm are shown in red boxes after pruning candidates of an extreme size or aspect ratio.

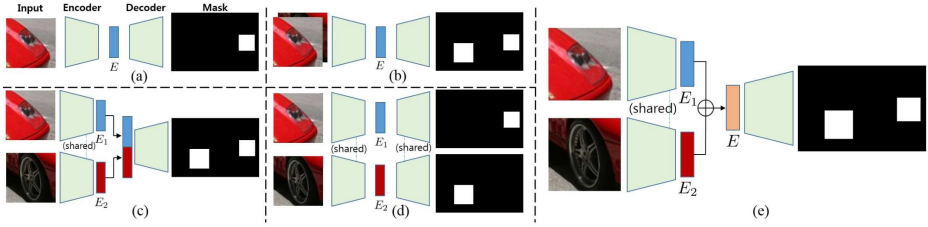

Fig. 4. Different structures of networks to predict a mask from input patches. We choose (e) as our encoder-decoder model.

3.1 Key Part Detection

We define key patches as informative local regions to generate the entire image. For example, when generating a face image, patches of eyes and a nose are more informative than those of the forehead and cheeks. Therefore, it would be better for the key patches to contain important parts that can describe objects in a target class. However, detecting such regions is a challenging problem as it requires to possess high-level concepts of the image. Although there exist methods to find most representative and discriminative regions [32,1], these schemes are limited to the detection or classification problems. In this paper, we only assume that key parts can be obtained based on the objectness score. The objectness score allows us to exclude most regions without textures or full of simple edges which unlikely contain key parts. In particular, we use the Edgebox algorithm [44] to detect key patches of general objects in an unsupervised manner. In addition, we discard detected patches with extreme sizes or aspect ratios. Figure 3 shows examples of detected key patches from various objects and scenes. Overall, the detected regions from these object classes are fairly informative. We sort candidate regions by the objectness score and feed the top N patches to the proposed network. In addition, the training images and corresponding key patches are augmented using a random left-right flip with the equal probability.

3.2 Part Encoding Network

The structure of the generator is based on the encoder-decoder network [6]. It uses convolutional layers as an encoder to reduce the dimension of the input data until the bottleneck layer. Then, transposed convolutional layers upsample the embedded vector to its original size. For the case with a single input, the network has a simple structure as shown in Figure 4(a). For the case with multiple inputs as considered in the proposed network, there are many possible structures. In this work, we carefully examine four cases while noting that our goal is to encode information invariant to the ordering of image patches.

The first network is shown in Figure 4(b), which uses depth-concatenation of multiple patches. This is a straightforward extension of the single input case. However, it is not suitable for the task considered in this work. Regardless of the order of input patches, the same mask should be generated when the patches have the same appearance. Therefore, the embedded vector E must be the same for all different orderings of inputs. Nevertheless, the concatenation causes the network to depend on the ordering, while key patches have an arbitrary order since they are sorted by the objectness score. In this case, the part encoding network cannot learn proper filters. The same issue arises in the model in Figure 4(c). On the other hand, there are different issues with the network in Figure 4(d). While it can resolve the ordering issue, it predicts a mask of each input independently, which is not desirable as we aim to predict masks jointly. The network should consider the appearance of both input patches to predict positions. To address the above issues, we propose to use the network in Figure 4(e). It encodes multiple patches based on a Siamese-style network and summarizes all results in a single descriptor by the summation, i.e., $E = E_1 + ... + E_N$. Due to the commutative property, we can predict a mask jointly, even if inputs have an arbitrary order. In addition to the final bottleneck layer, we use all convolutional feature maps in the part encoding network to construct U-net [27] style architectures as shown in Figure 2.

3.3 Mask Prediction Network

The U-net is an encoder-decoder network that has skip connections between i-th encoding layer and $(L-i)$-th decoding layer, where L is the total number of layers. It directly feeds the information from an encoding layer to its corresponding decoding layer. Therefore, combining the U-net and a generation network is effective when the input and output share the same semantic [7]. In this work, the shared semantic of input patches and the output mask is the target image.

We pose the mask prediction as a regression problem. Based on the embedded part vector E, we use transposed convolutional layers with a fractional stride [24] to upsample the data. The output mask has the same size as the target image and has a value between 0 and 1 at each pixel. Therefore, we use the sigmoid activation function at the last layer.

The spatial loss, $\mathcal{L}_S$, is defined as follows:

$$\mathcal{L}_S(G_{\mathcal{M}}) = \mathbb{E}_{\mathbf{x}\sim p_{data}(\mathbf{x}), M\sim p_{data}(M)}[\|G_{\mathcal{M}}(\mathbf{x}) - M\|_1]. \quad (2)$$

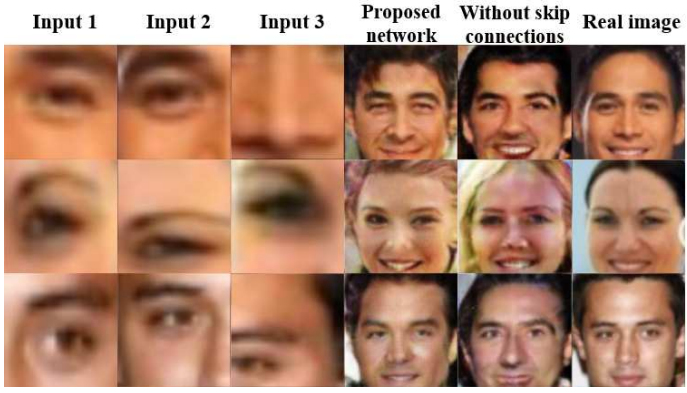

Fig. 5. Sample image generation results on the CelebA dataset using the network in Figure 2. Generated images are sharper and realistic with the skip connections.

We note that other types of losses, such as the l_2-norm, or more complicated network structures, such as GAN, have been evaluated for mask prediction, and similar results are achieved by these alternative options.

3.4 Image Generation Network

We propose a doubled U-net structure for the image generation task as shown in Figure 2. It has skip connections from both the part encoding network and mask generation network. In this way, the image generation network can communicate with other networks. This is critical since the generated image should consider the appearance and locations of input patches. Figure 5 shows generated images with and without the skip connections. It shows that the proposed network improves the quality of generated images. In addition, it helps to preserve the appearances of input patches. Based on the generated image and predicted mask, we define the appearance loss $\mathcal{L}_A$ as follows:

$$\mathcal{L}_A(G_\mathcal{M}, G_\mathcal{I}) = \mathbb{E}_{\mathbf{x},y\sim p_{data}(\mathbf{x},y),M\sim p_{data}(M)}[\|G_\mathcal{I}(\mathbf{x}) \otimes G_\mathcal{M}(\mathbf{x}) - y \otimes M\|_1], \quad (3)$$

where $\otimes$ is an element-wise product.

3.5 Real-Fake Discriminator Network

A simple discriminator can be trained to distinguish real images from fake images. However, it has been shown that a naive discriminator may cause artifacts [30] or network collapses during training [17]. To address this issue, we propose a new objective function as follows:

$$\begin{aligned}\mathcal{L}_R(G_\mathcal{I}, D) = {} & \mathbb{E}_{y\sim p_{data}(y)}[\log D(y)]+ \\ & \mathbb{E}_{\mathbf{x},y,y'\sim p_{data}(\mathbf{x},y,y'),M\sim p_{data}(M)} \\ & [\log(1 - D(G_\mathcal{I}(\mathbf{x})))+ \\ & \log(1 - D(M \otimes G_\mathcal{I}(\mathbf{x}) + (1 - M) \otimes y)) + \log(1 - D((1 - M) \otimes G_\mathcal{I}(\mathbf{x}) + M \otimes y))+ \\ & \log(1 - D(M \otimes y' + (1 - M) \otimes y)) + \log(1 - D((1 - M) \otimes y' + M \otimes y))],\end{aligned} \quad (4)$$

where y' is a real image randomly selected from the outside of the current mini-batch. When the real image y is combined with the generated image $G_{\mathcal{I}}(\mathbf{x})$ (line 4-5 in (4)), it should be treated as a fake image as it partially contains the fake image. When two different real images y and y' are combined (line 6-7 in (4)), it is also a fake image although both images are real. It not only enriches training data but also strengthens discriminator by feeding difficult examples.

4 Experiments

Experiments for the CelebA-HQ and CompCars datasets, images are resized to the have the minimum length of 256 pixels on the width or height. For other datasets, images are resized to 128 pixels. Then, key part candidates are obtained using the Edgebox algorithm [44]. We reject candidate boxes that are larger than 25% or smaller than 5% of the image size unless otherwise stated. After that, the non-maximum suppression is applied to remove candidates that are too close with each other. Finally, the image and top N candidates are resized to the target size, $256 \times 256 \times 3$ pixels for the CelebA-HQ and CompCars datasets or $64 \times 64 \times 3$ pixels for other datasets, and fed to the network. The λ_1 and λ_2 are decreased from 10^{-2} to 10^{-4} as the epoch increases. A detailed description of the proposed network structure is described in the supplementary material.

We train the network with a learning rate of 0.0002. As the epoch increases, we decrease λ_1 and λ_2 in (1). With this training strategy, the network focuses on predicting a mask in the beginning, while it becomes more important to generate realistic images in the end. The mini-batch size is 64, and the momentum of the Adam optimizer [9] is set to 0.5. During training, we first update the discriminator network and then update the generator network twice.

As this work introduces a new image generation problem, we carry out extensive experiments to demonstrate numerous potential applications and ablation studies as summarized in Table 1. Due to space limitation, we present some results in the supplementary material. All the source code and datasets will be made available to the public.

4.1 Datasets

The CelebA dataset [16] contains 202,599 celebrity images with large pose variations and background clutters (see Figure 8(a)). There are 10,177 identities with various attributes, such as eyeglasses, hat, and mustache.We use aligned and cropped face images of 108×108 pixels. The network is trained for 25 epochs.

Based on the CelebA dataset, we use the method [8] to generate a set of high-quality images. The CelebA-HQ dataset consists of 30,000 aligned images of $1,024 \times 1,024$ pixels for human face. The network is trained for 100 epochs.

There are two car datasets [37,11] used in this paper. The CompCars dataset [37] includes images from two scenarios: the web-nature and surveillance-nature (see Figure 8(c)). The web-nature data contains 136,726 images of 1,716 car models, and the surveillance-nature data contains 50,000 images. The network

Table 1. Setups for numerous experiments in this work.

Experiment	Description
Image generation from key patches	The main experiment of this paper. It aims to generate an entire image from key local patches without knowing their spatial location (Figure 6, Figure 8 and supplementary materials).
Image generation from random patches	It relaxes the assumption of the input from key patches to random patches. It is more difficult problem than the original task. We show reasonable results with this challenging condition.
Part combination	Generating images from patches of different objects. This is a new application of image synthesis as we can combine human faces or design new cars by a patch-level combination (Figure 9).
Unsupervised feature learning	We perform a classification task based on the feature representation of our trained network. As such, we can classify objects by only using their parts as an input.
An alternative objectvie function	It shows the effectiveness of the proposed objective function in (4) compared to the naive GAN loss. Generated images from our loss function is more realistic.
An alternative network structure	We evaluate three different network architectures; auto-encoder based approach, conditional GAN based method, and the proposed network without mask prediction network.
Different number of input patches	We change the number of input patches for the CelebA dataset. The proposed algorithm renders proper images for a different number of inputs.
Degraded input patches	To consider practical scenarios, we degrade the input patches using a noise. Experimental results demonstrate that the trained network is robust to a small amount of noise.
User study	As there is no rule of thumb to assess generated images, we carry out user study to evaluate the proposed algorithm quantitatively.

is trained for 50 epochs to generate 128×128 pixels images. To generate high-quality images (256×256 pixels), 30,000 training images are used and the network is trained for 300 epochs. The Stanford Cars dataset [11] contains 16,185 images of 196 classes of cars (see Figure 8(d)). They have different lighting conditions and camera angles. Furthermore, a wide range of colors and shapes, e.g., sedans, SUVs, convertibles, trucks, are included. The network is trained for 400 epochs.

The flower dataset [19] consists of 102 flower categories (see Figure 8(e)). There is a total of 8,189 images, and each class has between 40 and 258 images. The images contain large variations in the scale, pose, and lighting condition. We train the network for 800 epochs.

The waterfall dataset consists of 15,323 images taken from various viewpoints (see Figure 8(b)). It has different types of waterfalls as images are collected from the internet. It also includes other objects such as trees, rocks, sky, and ground, as images are obtained from natural scenes. For this dataset, we allow tall candidate boxes, in which the maximum height is 70% of the image height, to catch long water streams. The network is trained for 100 epochs.

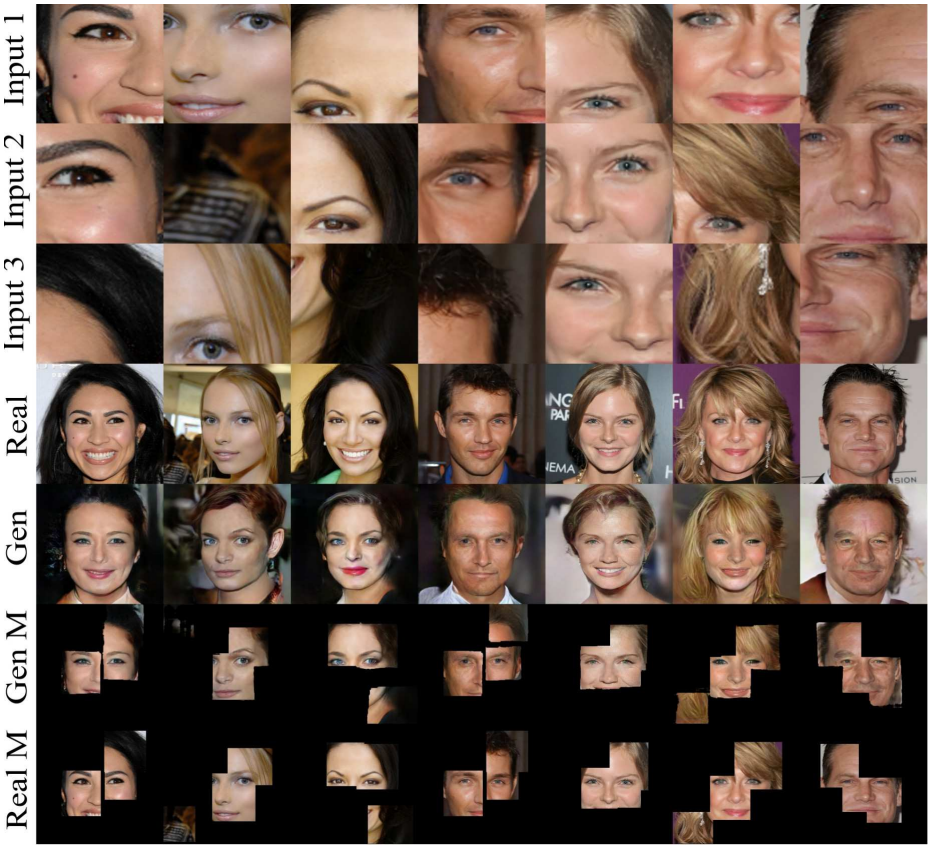

Fig. 6. Generated images and predicted masks on the CelebA-HQ dataset. Three key local patches (Input 1, Input 2, and Input 3) are from a real image (Real). Given inputs, images and masks are generated. We present masked generated images (Gen M) and masked ground truth images (Real M).

The ceramic dataset is made up of 9,311 side-view images (see Figure 8(f)). Images of both Eastern-style and Western-style potteries are collected from the internet. The network is trained for 800 epochs.

4.2 Image Generation Results

Figure 6, Figure 7, and Figure 8 shows image generation results of different object classes. Each input has three key patches from a real image and we show both generated and original ones for visual comparisons. For all datasets, which contain challenging objects and scenes, the proposed algorithm is able to generate realistic images. Figure 6 and Figure 7 show that the proposed algorithm is able to generate high-resolution images. In addition, input patches are well preserved around their original locations. As shown in the masked images, the

Fig. 7. Generated images and predicted masks on the CompCars dataset.

proposed problem is a superset of the image inpainting task since known regions are assumed to available in the latter task. While the CelebA-HQ dataset provides high-quality images, we can generate more diverse results on the original CelebA dataset as shown in Figure 8(a). The subject of the generated face images may have different gender (column 1 and 2), wear a new beanie or sunglasses (column 3 and 4), and become older, chubby, and with new hairstyles (column 5-8). Even when the input key patches are concentrated on the left or right sides, the proposed algorithm can generate realistic images (column 9 and 10). In the CompCars dataset, the shape of car images is mainly generated based on the direction of tire wheels, head lights, and windows. As shown in Figure 7 and Figure 8(c), the proposed algorithm can generate various poses and colors of cars while keeping the original patches properly. For some cases, such as column 2 in Figure 8(c), input patches can be from both left or right directions and the generation results can be flipped. It demonstrates that the proposed algo-

(a) CelebA dataset (b) Waterfall dataset

(c) CompCars dataset(128×128 pixels) (d) Stanford Cars dataset

(e) Flower dataset (f) Ceramic dataset

Fig. 8. Examples of generated masks and images on six datasets.

rithm is flexible since the correspondence between the generated mask and input patches, e.g., the left part of the mask corresponds to the left wheel patch, is not needed. Due to the small number of training samples compared to the CompCars dataset, the results of the Stanford Cars dataset are less sharp but still realistic. For the waterfall dataset, the network learns how to draw a new water stream (column 1), a spray from the waterfall (column 3), or other objects such as rock, grass, and puddles (column 10). In addition, the proposed algorithm can help restoring broken pieces of ceramics found in ancient ruins (see Figure 8(f)).

Figure 9 shows generated images and masks when input patches are obtained from different persons. The results show that the proposed algorithm can handle a wide scope of input patch variations. For example, inputs contain different skin colors in the first column. In this case, it is not desirable to exactly preserve inputs since it will generate a face image with two different skin colors. The proposed algorithm generates an image with a reasonable skin color as well as the overall shape. Other cases include with or without sunglasses (column

Fig. 9. Results on the CelebA dataset when input patches come from other images. Input 1 and Input 2 are patches from Real 1. Input 3 is a local region of Real 2. Given inputs, the proposed algorithm generates the image (Gen) and mask (Gen M).

2), different skin textures (column 3), hairstyle variations (column 4 and 5), and various expressions and orientations. Despite large variations, the proposed algorithm is able to generate realistic images.

5 Conclusions

We introduce a new problem of generating images based on local patches without geometric priors. Local patches are obtained using the objectness score to retain informative parts of the target image in an unsupervised manner. We propose a generative network to render realistic images from local patches. The part encoding network embeds multiple input patches using a Siamese-style convolutional neural network. Transposed convolutional layers with skip connections from the encoding network are used to predict a mask and generate an image. The discriminator network aims to classify the generated image and the real image. The whole network is trained using the spatial, appearance, and adversarial losses. Extensive experiments show that the proposed network generates realistic images of challenging objects and scenes. As humans can visualize a whole scene with a few visual cues, the proposed network can generate realistic images based on given unordered image patches.

Acknowledgements

The work of D. Lee, S. Choi, H. Yoo, and S. Oh is supported in part by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT (NRF-2017R1A2B2006136) and by 'The Cross-Ministry Giga KOREA Project' grant funded by the Korea government(MSIT) (No.GK18P0300, Real-time 4D reconstruction of dynamic objects for ultra-realistic service). The work of M.-H. Yang is supported in part by the National Natural Science Foundation of China under Grant #61771288, the NSF CAREER Grant #1149783, and gifts from Adobe and Nvidia.

References

1. Bansal, A., Shrivastava, A., Doersch, C., Gupta, A.: Mid-level elements for object detection. arXiv preprint arXiv:1504.07284 (2015) 6
2. Doersch, C., Gupta, A., Efros, A.A.: Unsupervised visual representation learning by context prediction. In: Proc. of the IEEE International Conference on Computer Vision (2015) 4
3. Dosovitskiy, A., Tobias Springenberg, J., Brox, T.: Learning to generate chairs with convolutional neural networks. In: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (2015) 3
4. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in Neural Information Processing Systems (2014) 2, 3, 5
5. Gregor, K., Danihelka, I., Graves, A., Rezende, D.J., Wierstra, D.: DRAW: A recurrent neural network for image generation. In: Proc. of the International Conference on Machine Learning (2015) 3
6. Hinton, G.E., Salakhutdinov, R.R.: Reducing the dimensionality of data with neural networks. Science **313**(5786), 504–507 (2006) 7
7. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (2017) 1, 4, 7
8. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of GANs for improved quality, stability, and variation. In: Proc. of the International Conference on Learning Representations (2018) 9
9. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. In: Proc. of the International Conference on Learning Representations (2014) 9
10. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013) 3
11. Krause, J., Stark, M., Deng, J., Fei-Fei, L.: 3D object representations for fine-grained categorization. In: Proc. of the IEEE International Conference on Computer Vision Workshops (2013) 6, 9, 10
12. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems (2012) 3
13. Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., Shi, W.: Photo-realistic single image super-resolution using a generative adversarial network. In: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (2017) 1, 4
14. Li, C., Wand, M.: Precomputed real-time texture synthesis with markovian generative adversarial networks. In: Proc. of the European Conference on Computer Vision (2016) 4
15. Liu, M.Y., Tuzel, O.: Coupled generative adversarial networks. In: Advances in Neural Information Processing Systems (2016) 1, 3
16. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: Proc. of International Conference on Computer Vision (2015) 6, 9
17. Metz, L., Poole, B., Pfau, D., Sohl-Dickstein, J.: Unrolled generative adversarial networks. Proc. of the International Conference on Learning Representations (2017) 8
18. Mirza, M., Osindero, S.: Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784 (2014) 3

19. Nilsback, M.E., Zisserman, A.: Automated flower classification over a large number of classes. In: Proc. of the IEEE Conference on Computer Vision, Graphics & Image Processing (2008) 6, 10
20. Noroozi, M., Favaro, P.: Unsupervised learning of visual representations by solving jigsaw puzzles. In: Proc. of the European Conference on Computer Vision (2016) 4
21. Oord, A.v.d., Kalchbrenner, N., Espeholt, L., Vinyals, O., Graves, A., et al.: Conditional image generation with pixelcnn decoders. In: Advances in Neural Information Processing Systems (2016) 3
22. Oord, A.v.d., Kalchbrenner, N., Kavukcuoglu, K.: Pixel recurrent neural networks. In: Proc. of the International Conference on Machine Learning (2016) 3
23. Pathak, D., Krähenbühl, P., Donahue, J., Darrell, T., Efros, A.A.: Context encoders: Feature learning by inpainting. In: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (2016) 4
24. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434 (2015) 3, 7
25. Reed, S., Akata, Z., Mohan, S., Tenka, S., Schiele, B., Lee, H.: Learning what and where to draw. In: Advances In Neural Information Processing Systems (2016) 3, 4
26. Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., Lee, H.: Generative adversarial text to image synthesis. In: Proc. of the International Conference on Machine Learning (2016) 3
27. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Proc. of the International Conference on Medical Image Computing and Computer-Assisted Intervention (2015) 4, 7
28. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training gans. In: Advances in Neural Information Processing Systems (2016) 3
29. Schawinski, K., Zhang, C., Zhang, H., Fowler, L., Santhanam, G.K.: Generative adversarial networks recover features in astrophysical images of galaxies beyond the deconvolution limit. Monthly Notices of the Royal Astronomical Society: Letters **467**(1), L110–L114 (2017) 1
30. Shrivastava, A., Pfister, T., Tuzel, O., Susskind, J., Wang, W., Webb, R.: Learning from simulated and unsupervised images through adversarial training. In: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (2017) 4, 8
31. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014) 3
32. Singh, S., Gupta, A., Efros, A.A.: Unsupervised discovery of mid-level discriminative patches. In: Proc. of the European Conference on Computer Vision (2012) 6
33. Vondrick, C., Pirsiavash, H., Torralba, A.: Generating videos with scene dynamics. In: Advances In Neural Information Processing Systems (2016) 1
34. Wang, X., Gupta, A.: Generative image modeling using style and structure adversarial networks. In: Proc. of the European Conference on Computer Vision (2016) 3
35. Weinzaepfel, P., Jégou, H., Pérez, P.: Reconstructing an image from its local descriptors. In: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (2011) 4

36. Wu, J., Zhang, C., Xue, T., Freeman, B., Tenenbaum, J.: Learning a probabilistic latent space of object shapes via 3D generative-adversarial modeling. In: Advances in Neural Information Processing Systems (2016) 1
37. Yang, L., Luo, P., Change Loy, C., Tang, X.: A large-scale car dataset for fine-grained categorization and verification. In: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (2015) 9
38. Yeh, R.A., Chen, C., Lim, T.Y., Schwing, A.G., Hasegawa-Johnson, M., Do, M.N.: Semantic image inpainting with deep generative models. In: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (2017) 1, 4
39. Yoo, D., Kim, N., Park, S., Paek, A.S., Kweon, I.S.: Pixel-level domain transfer. In: Proc. of the European Conference on Computer Vision (2016) 1, 4
40. Zhang, H., Xu, T., Li, H., Zhang, S., Huang, X., Wang, X., Metaxas, D.: Stack-GAN: Text to photo-realistic image synthesis with stacked generative adversarial networks. In: Proc. of the IEEE International Conference on Computer Vision (2017) 3
41. Zhang, Y., Xiao, J., Hays, J., Tan, P.: Framebreak: Dramatic image extrapolation by guided shift-maps. In: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (2013) 4
42. Zhao, J., Mathieu, M., LeCun, Y.: Energy-based generative adversarial network. In: Proc. of the International Conference on Learning Representations (2017) 3
43. Zhu, J.Y., Krähenbühl, P., Shechtman, E., Efros, A.A.: Generative visual manipulation on the natural image manifold. In: Proc. of the European Conference on Computer Vision (2016) 1, 4
44. Zitnick, C.L., Dollár, P.: Edge boxes: Locating object proposals from edges. In: Proc. of the European Conference on Computer Vision (2014) 6, 9