Magic Pencil: Generalized Sketch Inversion via Generative Adversarial Nets

Hua Zhang, Xiaochun Cao*

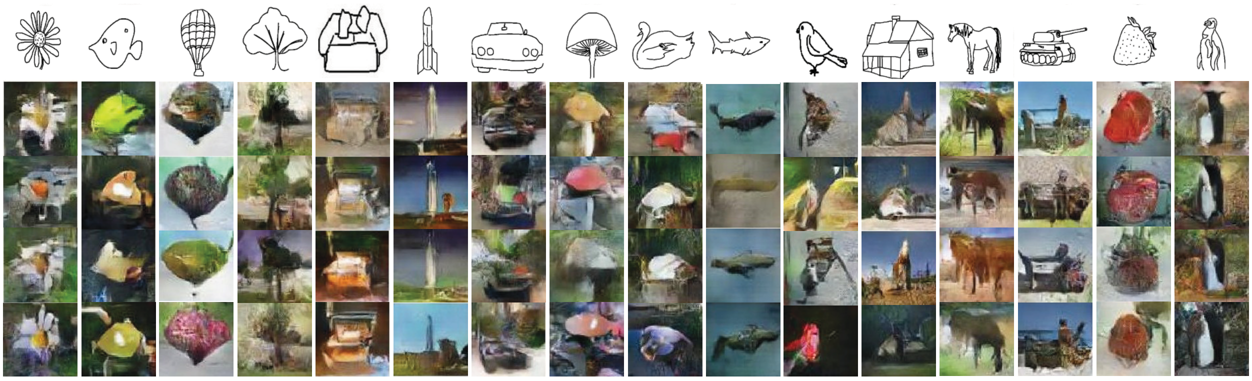

Figure 1: *The synthetic results based on our proposed method. The first row shows the sketch input, while the last four rows are the synthetic images according to the inputs. We can find that the generated real images have the consistency category information with the inputs. The categories of the sketches from left to right are: flower, fish, hot-air balloon, tree, couch, rocket, sedan, mushroom, swan, shark, songbird, cabin, horse, tank, strawberry, penguin.*

Abstract

Human can draw the sketches to show their imagination, which are more specific and vivid comparing with the text. The traditional work on sketch are focusing on recognition, which could be seen as the dimension reduction mapping from the sketch images to the semantic space. Different from the existing work, in this paper, we propose a novel image synthetic approach named **Magic Pencil**, which aims to synthesize the real images based on the sketch. Specifically, we design a novel multi-task neural network based on generative adversarial nets composed of three components i.e. the generator, the discriminator, and the classifiers. The generator is used to generate the real images after we input the sketch image and the variational factors. While the discriminator is designed to judge the reality of the generated images. And the classifiers is created to validate the semantic consistency between the sketch input and the generated images. We show that our proposed multi-task neural network trained end-to-end achieves promised results without further processing. To train our proposed neural network, we collect a novel dataset based on the existing sketchy dataset. From the experimental results, we could observe that our method could generate the semantic consistency and nearly reality images.

Keywords: Sketch inversion, generative adversarial net, deep neural network

Concepts: •**Computing methodologies** → **Image-based rendering;** •**Human-centered computing** → *Visualization techniques;*

1 Introduction

Hand-drawn sketches is the abstract representation of the objects in the real images, which preserves the essential shape information of the objects. Existing work on the sketch [Schneider and Tuytelaars 2014; Su et al. 2015] focus on sketch recognition and retrieval, which directly extract feature representation from the sketch images. However, it is a challenge work for the computer to automatically recognize the sketch images. The main challenge is that the sketch images contain the limited recognizable features i.e. the shape, little textures. Inspired by recent successful deep convolutional neural network [Chang et al. 2015; Li et al. 2015; Zhang et al. 2016; Yu et al. 2016], in this paper, we propose a generalized framework to inverse the sketch to real images while preserving their category information consistency.

To that end, we adopt the deep convlutional generative adversarial net [Goodfellow et al. 2014; Reed et al. 2016; Zhu et al. 2016], which has been demonstrated its ability to generate the real image from different modalities. Different from the traditional framework of GAN, the input of our proposed neural network including the sketch images and the variational factors instead of only noise vectors. Then, two inputs are fed into our designed neural network to generate the real images. To guarantee the reality of the generated images and the semantic consistency, we introduce a multi-task loss function to train the neural network. The generated real images are shown in Figure 1. The primary two contributions of our work are that:(1) we propose a novel deep neural network for sketch inversion; (2) Our proposed framework could inverse different categories of sketch images via the multi-task loss function.

2 Generative Adversarial Nets

Generative Adversarial Networks (GAN) is composed of two competition models, which are a generative model and a discriminative model. The generative model (G) is proposed to captures the data distribution of training images and generate the real images. While the discriminative model (D) is designed to penalize flaws in the generative model. Given a noise $z \sim p(z)$ and the real data

*e-mail:zhanghua, caoxiaochun@iie.ac.cn

SA '16, December 05-08, 2016, , Macao
ISBN: 978-1-4503-4540-8/16/12
DOI: http://dx.doi.org/10.1145/3005274.3005291

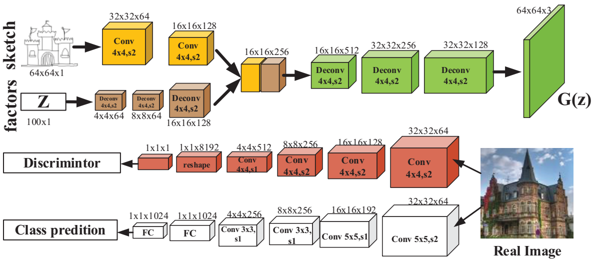

Figure 2: *Our proposed framework includes three components: the generator (top), the discriminator (middle), and the classifier (bottom). (please zoom in for the specific parameters) We use the deconvolution operator (Deconv) to upsample the feature maps, while use the convolution operators (Conv) to downsample the feature maps. The stride is abbreviate to s.*

$x \sim p(x)$, The objective function of GAN is the minimax game:

$$\min_{\theta_G} \max_{\theta_D} E_{x \sim p(x)}[log(D(x))] + E_{z \sim p(z)}[log(1 - D(G(z)))], \quad (1)$$

where θ_G, θ_D are the parameters of G and D, respectively. In [Goodfellow et al. 2014], the authors have proved that this objective function can achieve the global optimium precisely. In a word, the generator G is trained to fool D into misclassifying the fake samples, while D tries to distinguish the real images from synthetic images.

3 Technical Approach

Our proposed neural network is composed of three deep neural network, which are the Generator (G), the Discriminator (D), and the Classifiers (C). The basic parameters of our proposed neural network are listed in Figure 2. Given each sketch image, we firstly randomly generate a factor vector. Then the inputs are fed into the generator neural network to achieve the real images. Then, the faked images $G(z, sketch)$ is treated as the input of discriminator and classifiers. To update the parameters of the generator, we use the loss of the discriminator and the classifier via gradient back propagation. While we use the real images to update the parameters of the discriminator and the classifier.

The architecture of our neural network is shown in Figure 2, which contains the convolution layer, the pooling layer, the deconvolution layer and the fully connected layer. Specifically, the convolution and the pooling layer are used to downsample the input. While the deconvolution layer is designed to upsample the inputs. We use the ReLu as the activation function after each convolution and the deconvolution layers.

The loss function of our proposed neural network contains two components, and the optimization method is the Adam via the back propagation. The objective functions are:

$$\mathbf{L}_d = \sum_{i=1}^{M/2} log(D(x_i)) + \sum_{i=M/2+1}^{M} log(1 - D(G(sketch, z))), (2)$$

$$\mathbf{L}_c = -logP(y^i = k|x_i) = -log\frac{e^{-f^k(x_i)}}{\sum_{l=1}^{C} e^{-f^l(x_i)}}, \quad (3)$$

$$\mathbf{L}_{joint} = \mathbf{L}_d + \lambda \mathbf{L}_c, \quad (4)$$

where $\mathbf{L}_d$ and $\mathbf{L}_c$ is the loss function for the discriminator and the classifiers, respectively. x_i denotes the real images, and $G(sketch, z)$ is the synthetic images. C represents the number of categories and M is the number of total images including real images and the synthetic ones. λ is the weighting parameters to balance the components of loss function.

For the experimental setting, we use the sketchy dataset [Sangkloy et al. 2016], which includes the sketch image and the corresponding real image.There are 125 object categories in this dataset, and each sketch is drawn by the person after watching the real images. From the experiment results, we firstly find that the dimension of variation factors (z) influencing the quality of generated images. Different categories show the distinct generating capability with distinct dimension of z. Secondly, the loss of the classifier is also the factor influencing the results of the generator. Finally, a large number of real images could help for training a better generator.

Acknowledgements

This work was supported by National Key Research and Development Plan(No.2016YFC0800102), National Natural Science Foundation of China (No.61602464,61422213), "Strategic Priority Research Program" of the Chinese Academy of Sciences (XDA06010701), and National Program for Support of Top-notch Young Professionals.

References

CHANG, A. X., FUNKHOUSER, T., GUIBAS, L., HANRAHAN, P., HUANG, Q., LI, Z., SAVARESE, S., SAVVA, M., SONG, S., SU, H., ET AL. 2015. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*.

GOODFELLOW, I., POUGET-ABADIE, J., MIRZA, M., XU, B., WARDE-FARLEY, D., OZAIR, S., COURVILLE, A., AND BENGIO, Y. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, 2672–2680.

LI, Y., SU, H., QI, C. R., FISH, N., COHEN-OR, D., AND GUIBAS, L. J. 2015. Joint embeddings of shapes and images via cnn image purification. *ACM Trans. Graph 34*, 6, 234:1–234:12.

REED, S., AKATA, Z., YAN, X., LOGESWARAN, L., SCHIELE, B., AND LEE, H. 2016. Generative adversarial text to image synthesis. *arXiv preprint arXiv:1605.05396*.

SANGKLOY, P., BURNELL, N., HAM, C., AND HAYS, J. 2016. The sketchy database: learning to retrieve badly drawn bunnies. *ACM Transactions on Graphics (TOG) 35*, 4, 119.

SCHNEIDER, R. G., AND TUYTELAARS, T. 2014. Sketch classification and classification-driven analysis using fisher vectors. *ACM Transactions on Graphics (TOG) 33*, 6, 174.

SU, H., MAJI, S., KALOGERAKIS, E., AND LEARNED-MILLER, E. 2015. Multi-view convolutional neural networks for 3d shape recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, 945–953.

YU, Q., YANG, Y., LIU, F., SONG, Y.-Z., XIANG, T., AND HOSPEDALES, T. M. 2016. Sketch-a-net: A deep neural network that beats humans. *International Journal of Computer Vision*, 1–15.

ZHANG, H., LIU, S., ZHANG, C., REN, W., WANG, R., AND CAO, X. 2016. Sketchnet: Sketch classification with web images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1105–1113.

ZHU, J.-Y., KRÄHENBÜHL, P., SHECHTMAN, E., AND EFROS, A. A. 2016. Generative visual manipulation on the natural image manifold. In *Proceedings of European Conference on Computer Vision (ECCV)*.