영상변형:얼굴 스케치와 사진간의 증명가능한 영상변형 네트워크

숭타이리엥*, 이효종*
*전북대학교 컴퓨터공학
e-mail: hlee@chonbuk.ac.kr

Image Translation: Verifiable Image Transformation Networks for Face Sketch-Photo and Photo-Sketch

Thai-Leang Sung*, Hyo-Jong Lee*
*Division of Computer Science and Engineering, Chonbuk National University

Abstract

In this paper, we propose a verifiable image transformation networks to transform face sketch to photo and vice versa. Face sketch-photo is very popular in computer vision applications. It has been used in some specific official departments such as law enforcement and digital entertainment. There are several existing face sketch-photo synthesizing methods that use feed-forward convolution neural networks; however, it is hard to assure whether the results of the methods are well mapped by depending only on loss values or accuracy results alone. In our approach, we use two Resnet encoder-decoder networks as image transformation networks. One is for sketch-photo and another is for photo-sketch. They depend on each other to verify their output results during training. For example, using photo-sketch transformation networks to verify the photo result of sketch-photo by inputting the result to the photo-sketch transformation networks and find loss between the reversed transformed result with ground-truth sketch. Likely, we can verify the sketch result as well in a reverse way. Our networks contain two loss functions such as sketch-photo loss and photo-sketch loss for the basic transformation stages and the other two-loss functions such as sketch-photo verification loss and photo-sketch verification loss for the verification stages. Our experiment results on CUFS dataset achieve reasonable results compared with the state-of-the-art approaches.

1. Introduction

Sketch-photo generation is one of the image-to-image translation problems. It refers to a constrained synthesized task of transforming a sketch image to a real photo. Sketch-photo generation has been developed widely as applications in the field of computer vision and image processing. For example, several types of sketch-photo applications have been used in the police department and law enforcement where require the eyewitnesses or victims to draw sketch images of criminals.

Several convolutional neural networks [1-3] have been used to solve the problem of sketch-photo synthesis by mapping input image to the transformed synthesized image and penalizing the discrepancy between the synthesized image and ground-truth image using pixel loss functions such as L1, and L2. Then there are some GANs-based approaches like cGANs [4] also works on image-to-image translation including sketch or photo generation. However, all those approaches resulted in a limited manner such as blurry image, overlapped generated and detail losses.

Our paper proposes two identical image transformation networks with a reversal purpose – they are verification network for each other which meanwhile they are serving their own purposes. For example, one network is for sketch-photo synthesis and another one is for photo-sketch synthesis. So, when one input goes under one network the transformed input must be input to another network for the purpose of verification. Below is our transformation steps:

1) Sketch → Net I → Generated photo → Net II → Reversed-sketch.
2) Photo → Net II → Generated sketch → Net I → Reversed-photo.

We verified the reversal sketch and photo with the original sketch and photo using structural similarity index (SSIM).

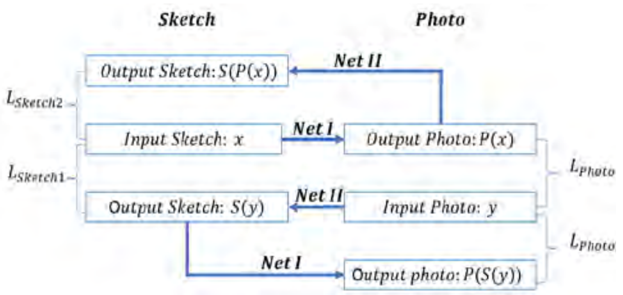
Both steps are not restricted in the order procedure. We can start from step 2 and do step 1 later. We also propose joint-loss functions including sketch loss, photo loss, sketch-verification loss, and photo verification loss and minimized them during training through backpropagation.

In summary, we contribute a new approach that allows us to transform sketch to photo and photo to sketch where we can achieve the qualitative images that sharper and realistic. We obtain two models after training which can be utilized separately in the real-world applications. The framework is built not restricted for sketch-photo/photo-sketch only but image-to-image translation as well. Our results are reasonable due to qualitative and quantitative evaluation.

We will explain our methods in Section 2 and show our experimental setup and results in Section 3. Finally, we will conclude our paper in Section 4.

2. Methods

In this section, we will explain our method in detail. Firstly, we will explain about our proposed architecture. Then we will explain our loss function we will use in this paper.



(Figure 1) Proposed verifiable image transformation networks for sketch-photo and photo-sketch

As shown in Figure 1, we applied our framework model to Sketch and Photo tasks. There have two transformation networks, Net I and Net II. In this study, we used Resnet encoder and decoder [5]. Net I is a Resnet image transformation network that encodes an input sketch image using encoder and synthesizes a photo image through decoder. Net II also has the same idea as Net I, but we use this network to synthesize sketch image from input photo.

<Table I> Resnet Encoder-Decoder

Layer	Output shape	Parameters
Input Image		
Conv Layer 1	[64, 256, 256]	9,472
Conv Layer 2	[128, 128, 128]	73,856
Conv Layer 3	[256, 64, 64]	295,168
Transformation	Resnet Block 1 Resnet Block 2 . . . Resnet Block 9	590,080 x 18
DeConv Layer 1	[128, 128, 128]	295,040
DeConv Layer 2	[64, 256, 256]	73,792
Conv Layer 3	[3, 256, 256]	9,411
Output Image		Total: 11,378,179

Based on the framework in Figure 2 at first, the framework starts training from inputting sketch x through Net I and obtains output photo P(x). The first photo loss L_{Photo1} is employed. This loss finds the distance between generated photo P(x) and ground-truth photo y. Simultaneously, it inputs photo image y through Net II and obtains sketch image S(y). Then the first sketch loss $L_{Sketch1}$ is employed. Then for the second step, the framework starts to verify the synthesized sketch S(y) and photo P(x). S(y) is input into Net I to gain a reversal photo P(S(y)) of the synthesized sketch S(y). Second photo loss function L_{Photo2} between P(S(y) and ground-truth y can be found here. We can also call this loss "sketch-photo verification loss". Reversely, P(x) is input into Net II to obtain a reversal sketch S(P(x)) of synthesized photo P(x). Second sketch loss function $L_{Sketch2}$ is found. We call it "photo-sketch verification loss" as well. The loss functions mentioned above are minimized during the training process through the networks' backpropagation iteratively.

The methods we use in this paper are the loss functions that we described above. We use a pixel-wise L1 loss function for sketch loss and photo loss and Structural Similarity Index (SSIM) [6] for sketch and photo verification loss. SSIM is a perceptual metric for predicting the perceived quality of an image as well as digital image and video. Different from a pixel-wise function such as L1 or L2, SSIM perceives structural information of visual scene's object, luminance and contrast masking terms. Therefore, it is a good option for the verification of image-to-image translation. The following is how we use SSIM as loss functions in our paper.

$$SSIM(p,t) = [l(p,t) \cdot c(p,t) \cdot s(p,t)], \qquad (1)$$

where p is a predicted image and t is the ground-truth image. SSIM is based on three comparison measurements between p and t: luminance l, contrast c, and structure s. Read [6] for how to calculate these three terms.

SSIM gives a maximum value 1 to the image that has the best quality. Therefore, we can define our loss function using SSIM as following:

$$Loss = 1 - SSIM(p,t). \qquad (2)$$

According to (1) and (2), we can write the four loss of functions of our framework below:

- Sketch loss:

  $$L_{Sketch1} = |x - S(y)|, \qquad (3)$$

  where x is the original sketch used as ground-truth, and S(y) is a generated sketch.
- Photo loss:

  $$L_{Photo1} = |y - P(x)|, \qquad (4)$$

  where y is the original photo used as ground-truth, and P(x) is a generated photo.
- Photo-sketch verification loss:

  $$L_{Sketch2} = 1 - SSIM\big(S\big(P(x)\big), x\big), \qquad (5)$$

  where S(P(x)) is a reversal generated sketch of the generated photo.
- Sketch-photo verification loss:

  $$L_{Photo2} = 1 - SSIM\big(P\big(S(y)\big), y\big). \qquad (6)$$

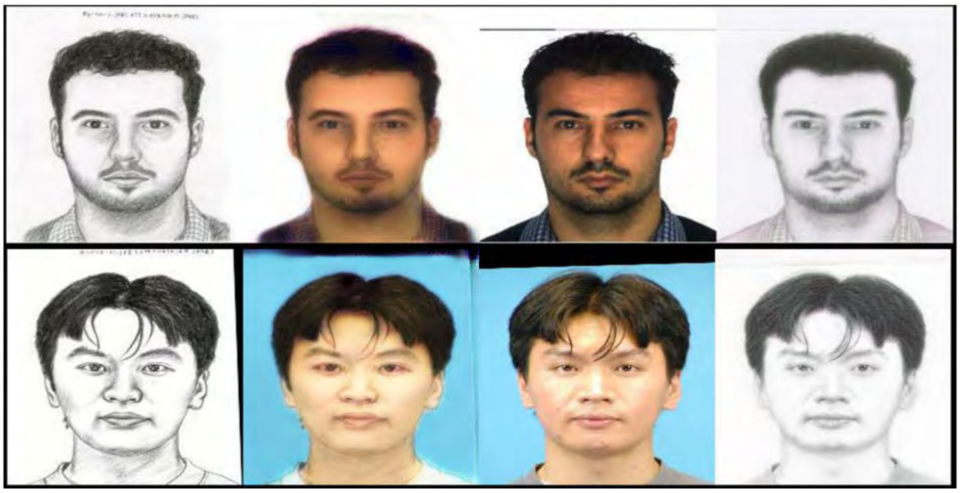
  where P(S(y)) is a reversal generated photo of the generated sketch.

$$Total\ loss = \frac{(L_{Sketch1} + L_{Sketch2} + L_{Photo2} + L_{Photo1})}{4}. \qquad (7)$$

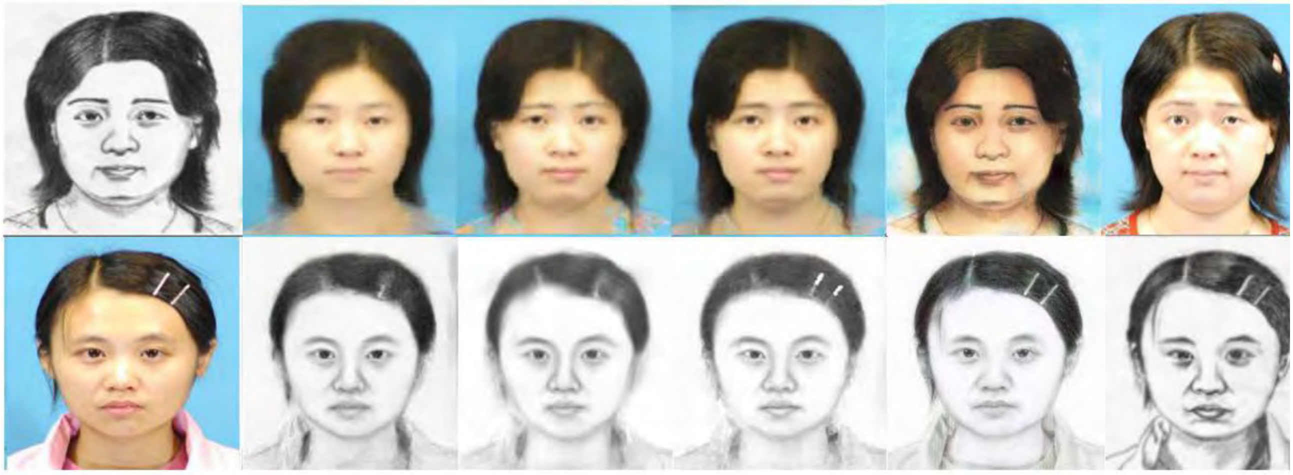
Again, our framework contains two Resnet encoder-decoder networks, Net I for sketch-photo and Net II for photo-sketch transformation. Both networks have the same network architectures as <Table I>. In each convolution, layer consists of Convolution-Batchnorm-ReLu and at the last layer, we use Tanh activation function instead of ReLu.

3. Experimental Setup and Results

We evaluated our methods on CUFS dataset [7] for both transformations, sketch-photo and photo-sketch. We use 50% of the data for training and the other 50% for testing. Our experiments were trained on NVIDIA Titan X (Pascal) GPUs and used Pytorch. We set Adam Solver to 0.0002, Momentum



(Figure 2) Results from our verifiable image transformation network



(Figure 3) Comparison of our results with existing approaches

<Table II> Comparison with the existing state-of-the-arts

Photo-Sketch		
	SSIM	VIF
MWF	0.50	0.79
S-FSPS	0.62	**0.19**
Ours	**0.68**	0.18
Sketch-Photo		
	SSIM	VIF
MWF	0.35	**0.70**
S-FSPS	0.51	0.13
Ours	**0.87**	0.48

to 0.5 for our optimizers. We set batch size to 6 and training epochs at the maximum of 4000.

We conducted qualitative and quantitative experiments to evaluate the performance of the transformed images. We use SSIM, and VIF [9] as our evaluation metrics.

(Figure 2) is our results of transforming sketch to photo, and as well as a photo to sketch. We compare our results quality with some of the state-of-the-art such as MWF [10], transductive [11] method and S-FSPS [12] in (Figure 3). Our results are free from the overlapping of the image pattern and less blurred. Then, in <Table II> is the list of our quantitative result with these approaches. As shown in the table, our structural similarity scores are higher than the existing method for both sketch-photo and photo-sketch. However, we lost to S-FSPS on Visual Information Fidelity at Photo-sketch. We also lost to MWF at sketch-photo.

4. Conclusion

We proposed two image transformation networks to do the tasks of sketch-photo and photo-sketch. Both networks are verifiers for each other. The synthesized image still maintains its original patterns even though it has been transformed into another modality. Moreover, our networks refine themselves every time they have their verifications done during training. Our approach is two-birds-with-one-stone achievement, because, at the end of the day, we gain two models for two specific transformation tasks, respectively. Accordingly, our methods do not restrict on sketch and photo only. It is recommended to use this approach for the other image-to-image translation tasks as well.

Acknowledgment

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (GR 2016R1D1A3B03931911).

References

[1] X. Fu, J. Huang, X. Ding, Y. Liao, and J. Paisley, "Clearing the skies: A deep network architecture for single-image rain removal," *IEEE Transactions on Image Processing*, vol. 26, no. 6, pp. 2944-2956, 2017.

[2] R. Zhang, P. Isola, and A. A. Efros, "Colorful image colorization," in *European Conference on Computer Vision*, 2016, pp. 649-666: Springer.

[3] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2536-2544, 2016.

[4] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 5967-5976: IEEE.

[5] G. E. Hinton and R. R. J. s. Salakhutdinov, "Reducing the dimensionality of data with neural networks," vol. 313, no. 5786, pp. 504-507, 2006.

[6] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. J. I. t. o. i. p. Simoncelli, "Image quality assessment: from error visibility to structural similarity," vol. 13, no. 4, pp. 600-612, 2004.

[7] X. Wang, X. J. I. T. o. P. A. Tang, and M. Intelligence, "Face photo-sketch synthesis and recognition," vol. 31, no. 11, pp. 1955-1967, 2009.

[8] Z. Wang and A. C. J. I. s. p. l. Bovik, "A universal image quality index," vol. 9, no. 3, pp. 81-84, 2002.

[9] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Trans Image Process,* vol. 15, no. 2, pp. 430-44, Feb 2006.

[10] H. Zhou, Z. Kuang, and K.-Y. K. Wong, "Markov weight fields for face sketch synthesis," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 1091-1097: IEEE.

[11] D. T. N. Wang, X. Gao, X. Li, and J. Li "Transductive face sketchphoto synthesis," *IEEE Trans. Neural Netw. Learn. Syst.,* vol. vol. 24, no. 9, 2013.

[12] C. Peng, X. Gao, N. Wang, J. J. I. T. o. C. Li, and S. f. V. Technology, "Superpixel-based face sketch–photo synthesis," vol. 27, no. 2, pp. 288-299, 2017.