Unpaired Image-to-Sketch Translation Network for Sketch Synthesis

Yue Zhang, Guoyao Su, Yonggang Qi, Jie Yang
Beijing University of Posts and Telecommunications, Beijing, China

***Abstract*—Image-to-sketch translation is to learn the mapping between an image and a corresponding human drawn sketch. Machine can be trained to mimic the human drawing process using a training set of aligned image-sketch pairs. However, to collect such paired data is quite expensive or even unavailable for many cases since sketches exhibit various level of abstractness and drawing preferences. Hence we present an approach for learning an image-to-sketch translation network via unpaired examples. A translation network, which can translate the representation in image latent space to sketch domain, is trained in unsupervised setting. To prevent the problem of representation shifting in cross-domain translation, a novel cycle+ consistency loss is explored. Experimental results on sketch recognition and sketch-based image retrieval demonstrate the effectiveness of our approach.**

I. Introduction

Sketching has a long history in human society that people are able to draw a few line strokes to record visual world since ancient times. Defined as sketch synthesis in computer vision, which aims to teach machine to generate sketch from real image just as humans do, has been attracted increasing attentions lately. Human visual system is so powerful that people can easily draw a sketch to express a complex real-world object just given a glance, whereas it is quite challenging for machine to perform similar ability due to the inherent ambiguities in sketch, e.g. highly abstractness and large appearance variance [15], [17], thus leading to severe cross-domain gap between image and sketch [13], [14]. Recently, due to the success of generative adversarial learning [4], sketch synthesis could be treated as an image-to-image translation problem [7], [18]. However, almost all the prior arts [1], [14] typically require tens of thousands image-sketch paired training examples to alleviate the above mentioned difficulties. Requiring such a large amount of data is notorious since it is labour costly to collect the one-to-one mapping image-sketch paired data.

Therefore, in this paper, we propose an unsupervised image-to-sketch translation network which could be trained only given unpaired image-sketch data. The problem of unpaired/unsupervised image-to-sketch translation is difficult due to the large cross-domain gap—no paired examples showing how a real image could be transferred to a corresponding human sketch. Similar problem has been studied for unpaired image-to-image translation, which has achieved impressive results by using cycle consistency based on coupled GANs [6], [12], [18]. However, our unpaired image-to-sketch learning task is considered as much harder due to the larger domain gap exists comparing with the image-to-image case. To solve this problem, an end-to-end network based on variational autoencoder (VAE) [9] and generative adversarial network (GAN) is proposed. Specifically, we model each domain using VAE to obtain their encoder and decoder, i.e., (E_{image}, D_{image}) and (E_{sketch}, D_{sketch}). Then we attempt to learn a translation network (*TranNet*) to convert the representation in image domain to sketch domain, i.e., $T_{I\to S}(image) \to sketch$, which could be further used to generate a corresponding sketch $D_{sketch}(T_{I\to S}(image))$. In particular, a novel *Cycle+ Consistency* is developed to explicitly restrict the representations in two latent spaces for the same input image to be consistent. Edge of real image is importantly embedded as an additional shape prior for regulating the translation of the representations to prevent representation shifting, hence can enforce a better image-sketch resemblance in appearance.

The contributions of this paper can be summarized as follows: (i) an unsupervised model based on VAE-GAN is proposed for stroke-level sketch synthesis by using unpaired image-sketch data. (ii) A novel cycle+ consistency loss is designed for regulating the domain-specific representations to be consistent, hence restricting the *TranNet* to be instance sensitive. (iii) Edge cue is utilized to further constrain the *TranNet* to learn to encode shape knowledge provided by input image. (iv) Our model is also applicable to generate image from sketch in reverse order.

II. Related Work

Unpaired Image-to-Image Translation The problem of sketch synthesis can be categorized as image-to-image translation. A large body of literature on translation algorithms in the supervised setting [7], [8], [16], which assumes paired examples are available. However in many cases, requiring paired data is expensive. Hence several models are presented to tackle the unpaired setting lately. CycleGAN [18] adopts a bidirectional mapping model based on coupled GAN with cycle consistency loss. UNIT [12] assumes a shared-latent space existed across two domains and a framework based on coupled GANs is proposed. MUNIT [6], [3] and DRIT [11] further decompose the feature space into shared content space and domain specific style space, hence achieve better diversity and improved quality on produced image. However, most works tackle on pixel-to-pixel level, which cannot be readily applied to pixel-to-stroke translation problem.

Vector Sketch Generation Despite the success of pixel-level image generation on image synthesis and editing [7], generalizing them onto sketch is proven to be nonsense [14].

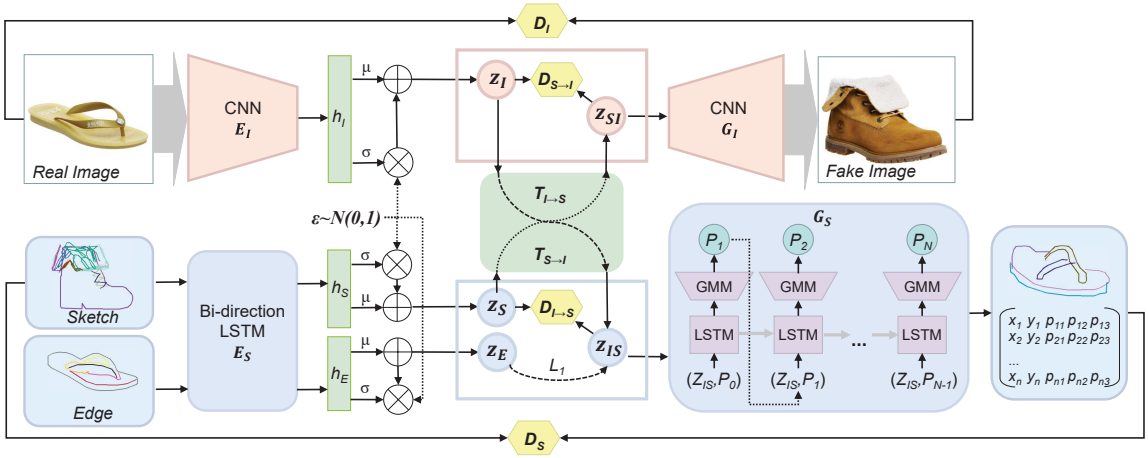

Fig. 1. Overview of network training. Note that edge is only used for training.

Learned from how humans draw object, sketch-rnn [5], which is proposed to express sketch as a sequential vector representation of strokes, opens the door for vector sketch generation. Specifically, a RNN-based VAE model is proposed to learn from pen stroke actions of humans, which can generate impressive sketch generation results. While it cannot be applied for cross-domain translation. Pix2seq [1] replaced sketch-rnn with a CNN encoder, hence can produce several novel sketches in vector format given an input sketch image. However, regarding the generated sketch details, the resemblance to the corresponding input image is unsatisfied [14]. In [14], shortcut cycle consistency loss is developed and a hybrid supervised-unsupervised multi-task learning framework is proposed for sketch synthesis. Although promising results can be obtained, a large number of paired image-sketch examples are required for training. On the contrary, ours is a fully unsupervised approach that could be trained without paired data.

III. Methodology

A. Model Overview

The goal is to learn the mapping between real image and stroke-level vector sketch in an unsupervised way. Given a set of unpaired real images $I \sim P_I$ and sketches in vector format $S \sim P_S$, we aim to (i) learn two encoders E_I and E_S to map image and sketch into their latent space, i.e., $E_I : I \to Z_I$ and $E_S : S \to Z_S$, thus (ii) to translate the representation in one domain to another by learning translation networks, $T_{I\to S} : Z_I \to Z_S$ for image-to-sketch, and $T_{S\to I} : Z_S \to Z_I$ in reverse order, and (iii) to learn two mappings $G_I : Z_I \to I$, $G_S : Z_S \to S$, which can generate image and sketch from their latent space representations respectively. In addition, adversarial discriminators including D_I, D_S, $D_{I\to S}$ and $D_{S\to I}$ are introduced, where D_I and D_S aim to distinguish between real and translated data in image and sketch domain respectively; $D_{I\to S}$ is to distinguish between representation in sketch domain Z_S and the translated representation in sketch domain from image domain Z_{IS}; Similarly, $D_{S\to I}$ is to discriminate between Z_I and Z_{SI}. Fig. 1 shows the training process of our network, details can be found in the following.

B. Objectives

VAE Loss We apply VAE losses both on image and sketch branches to learn their latent space separately. For image branch, E_I is a CNN encoder outputs a hidden feature h_I, which is further projected into μ_I and σ_I by using fully connected layer, hence to construct latent feature vector $z_I \sim N(\mu_I, \sigma_I^2)$ where $z_I \in Z_I$. z_I is then fed into the generator G_I which is a transposed CNN to reconstruct image. Hence the loss is defined as:

$$L_I^{VAE} = L_I^{recon} + \alpha L_I^{KL} \tag{1}$$

where $L_I^{recon} = \mathbb{E}_{i\sim P_I}\left[\|G_I(E_I(i)) - i\|_1\right]$ is the reconstruction loss, and $L_I^{KL} = KL\left(N(\mu,\sigma^2)||N(0,1)\right)$ is the KL divergence distance between $N(\mu,\sigma^2)$ and $N(0,1)$. α controls the relative importance of the KL loss.

For sketch branch, given a sketch described by a set of points, each point is denoted as (x, y, p_1, p_2, p_3), where (x, y) is the position and (p_1, p_2, p_3) denotes different pen actions. E_S is a sketch encoder built on bi-directional LSTM. Our generator G_S is a GMM embedded LSTM adopts a recurrent structure, which feed the previous estimated stroke point by GMM as input for the next prediction. This design suits for our unsupervised setting, since no paired sketch strokes are available at training stage, and facilitates sketch synthesis that could be achieved given an image only, $G_S(T_{I\to S}(Z_I))$. Similar to [14], the loss is defined as:

$$\begin{aligned} L_S^{VAE} = &-\frac{1}{N_{max}}\Big(\sum_{i=1}^{N_s}\log(P_{GMM}(x_i, y_i|\theta_i)) \\ &+ \sum_{i=1}^{N_{max}}\sum_{j=1}^{3} p_{ij}\log(q_{ij})\Big) \end{aligned} \tag{2}$$

where N_{max} represents the upper bound of the amount of stroke points in one sketch and N_s means the number of points actually exist in a sketch. p_{ij} and q_{ij} are the real and predicted distribution of painting action separately.

Adversarial Loss Adversarial losses are used to (i) enforce the generated data to be undistinguished from real data; For the image-to-sketch translation $F = \{E_I, T_{I\to S}, G_S\} : I \to S$

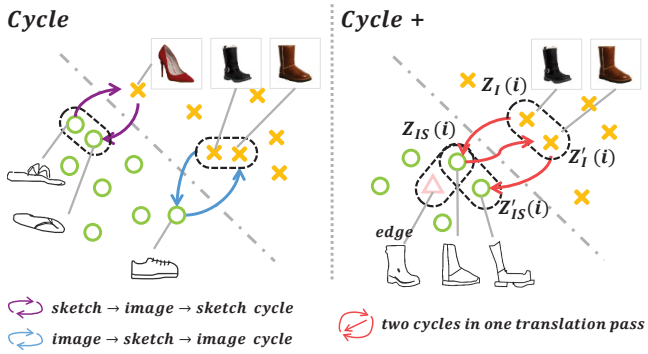

Fig. 2. Comparison between original cycle consistency and our proposed cycle+ consistency loss. Best viewed in color.

and its discriminator D_S, the adversarial loss is:

$$L_{I\to S}^{GAN} = \mathbb{E}_{s\sim P_S}[\log D_S(s)] + \mathbb{E}_{i\sim P_I}[\log(1 - D_S(F(i)))] \quad (3)$$

similarly, the adversarial loss for sketch-to-image translation is $L_{S\to I}^{GAN}$. (ii) further enforce the translated latent representation to be undistinguished from real ones. For image-to-sketch representation translation $F' = \{E_I, T_{I\to S}\} : Z_I \to Z_S$ and its discriminator $D_{I\to S}$, the loss can be defined as:

$$L_{Z_I\to Z_S}^{GAN} = \mathbb{E}_{s\sim S}[\log D_{I\to S}(Z_s)] + \mathbb{E}_{i\sim I}[\log(1 - D_{I\to S}(F'(i)))] \quad (4)$$

similar loss $L_{Z_S\to Z_I}^{GAN}$ could be also defined.

Cycle+ Consistency Loss As shown in Fig. 2, original cycle consistency [18] does not theoretically guarantee the appearance consistency between the synthesized sketch and the reference image in our problem. In other words, shift on latent representation might happen, which is non-trivial in the absence of unpaired examples that would lead to mismatching translation, e.g. a low-boot sketch generated from a high-boot image. To address this issue, the image i corresponding edge i_{edge} in vector format is exploited as auxiliary knowledge for explicitly regulating synthesized sketch to be consistent to input image, by pulling $z_{IS} = T_{I\to S}(E_I(i))$ together with $z_E = E_S(i_{edge})$ within the sketch latent space. Formally, this constraint is a L_1 loss, $L_{edge} = \mathbb{E}_{i\sim P_I}[\|z_{IS} - z_E\|_1]$. Additionally, we uniformly formulate the image-to-sketch cycle and sketch-to-image cycle in one translation pass, i.e., regard to the translation started from image i: $Z_I(i) \to Z_{IS}(i) \to Z_I^{'}(i) \to Z_{IS}^{'}(i)$ as shown in Fig 2. Hence the final cycle+ consistency loss is defined as:

$$L_{cycle+} = L_{edge} + L_{I\to S\to I} + L_{S\to I\to S} \quad (5)$$

where $L_{I\to S\to I} = \mathbb{E}_{i\sim P_I}[\|Z_I(i) - Z_I^{'}(i)\|_1]$ and $L_{S\to I\to S} = \mathbb{E}_{i\sim P_I}[\|Z_{IS}(i) - Z_{IS}^{'}(i)\|_1]$. *Note that edge is only used at training stage.*

Full Objective Our full objective is :

$$\begin{aligned} L_{full} = {} & L_{I\to S}^{GAN} + L_{S\to I}^{GAN} + L_{Z_I\to Z_S}^{GAN} + L_{Z_S\to Z_I}^{GAN} \\ & + \lambda(L_I^{VAE} + L_S^{VAE}) + \gamma L_{cycle+} \end{aligned} \quad (6)$$

where λ and γ are used to control the importance of VAE loss and cycle+ consistency loss. After training, sketch synthesis is achieved by applying image encoder E_I, translation network $T_{I\to S}$ and sketch decoder G_S given an input image i, i.e., $G_S(T_{I\to S}(E_I(i)))$.

IV. Experiments

Follow [14], experiments on sketch recognition and fine-grained SBIR are used for evaluating the quality of the synthesised sketch. *Shortcut Cycle* [14] and *Pix2seq* [1], which both work on paired examples, serve as alternatives for comparison since no existing unsupervised methods for image-to-sketch translation. We also show our model is capable of sketch-to-image translation in this section.

Experimental setting and datasets: QMUL-Shoe-Chair-V2 [17], which is the largest fine-grained image-sketch dataset, is utilized for evaluation. Specifically, the split of shoes, which contains 2000 images and 6648 sketches, are used in our experiments. The dataset is split into training and testing set by ratio of 9:1 that there are totally 1800 images with 5982 sketches for training and 200 images with 666 sketches for testing. To train our model, we randomly pair an image with a sketch in training set. Hence there are $1800 * 5982 \approx 10^7$ samples for training.

Competitors: *Shortcut Cycle* [14] and *Pix2seq* [1] are state-of-the-arts on sketch synthesis, which *Shortcut Cycle* is a hybrid supervised-unsupervised learning model and *Pix2seq* is a fully supervised method. More specifically, we retrain these two models by using the default training set split of QMUL-Shoe-Chair-V2 dataset for comparisons. In addition, to illustrate the effectiveness of key components, the alternative versions based on our full model are also evaluated, including our full model trained without translation network (*Full - TranNet*), without edge used (*Full - Edge*) and without cycle+ consistency loss (*Full - Cycle+*).

Evaluation Metrics The same metrics are used follow [14]: (i) *Recognition Score*: To evaluate the how recognizable of generated sketches, a CNN-based classifier [10] is trained on TU-Berlin dataset [2] which contains 20,000 sketches over 250 categories. Then it is used to test if the synthesised sketch could be correctly recognized as shoe in class level. (ii) *FG-SBIR Accuracy*: A FG-SBIR triplet network [17] is retrained on the QMUL-Shoe-Chair-V2 training set, to evaluate the appearance resemblance of the synthesised sketch and the input image.

Results and discussions Example qualitative results are shown in Fig. 3. We can observe that our full model can generate various types of shoe sketches, and the line drawings are simpler but more realistic with finer details comparing against with other competitors. It is interesting that the shoelace is even nicely drawn by our model (fifth and sixth row), while sketches synthesized by *Shortcut Cycle* and *Pix2Seq* often contains some unrealistic drawing shapes, lines and wrong details. Additionally, we can witness the importance of the key components of our model, the synthesised shoes can hardly resemble to the input image references regards in both overall shape and details. Quantitative results are shown in Table I. The recognition scores suggest that our full model achieves the best (65.84%) comparing against with *Shortcut* (59.78%) and *Pix2Seq* (31.61%) in acc.@1. Ours drop down to the second place in acc.@10 but still close to the best

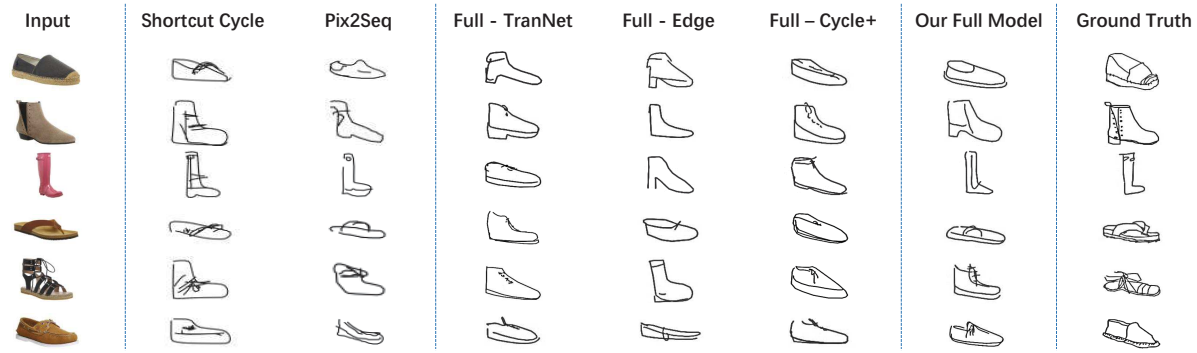

Fig. 3. Comparison of qualitative results.

TABLE I
QUANTITATIVE RESULTS IN TOP-1 AND TOP-10 ACCURACY OF RECOGNITION AND FG-SBIR.

	Recognition		FG-SBIR	
Method	***acc.@1***	***acc.@10***	***acc.@1***	***acc.@10***
Human sketch	78.02%	94.98%	9.73%	44.74%
Shortcut Cycle	59.78%	89.71%	**2.98%**	**17.43%**
Pix2Seq	31.61%	65.41%	2.04%	11.77%
Full - TranNet	83.45%	94.69%	0.63%	5.02%
Full - Edge	62.39%	82.98%	0.78%	6.91%
Full - Cycle+	**85.53%**	**95.92%**	1.26%	8.63%
Our Full Model	65.84%	85.68%	2.51%	17.27%

algorithm *Shortcut* (85.68% vs 89.71%). Interestingly, our *Full - Cycle+* and *Full - TranNet* largely outperforms the others and even is better than human data both in acc.@1 and acc.@10. This is because these two synthesizers can draw good shoes, which *Full - TranNet* is a pure variational encoder and *Full - Cycle+* gains powerful generative ability from adversarial learning, but both of them have clear drawbacks on the resemblance level between synthesized sketch and its input image (See Fig. 3). It is testified by the results of FG-SBIR accuracy that *Full - Cycle+* and *Full - TranNet* are far more worse than our full model and other competitors. The overall scores for FG-SBIR are quite low that even human sketch can only obtain about 10% in acc.@1 and 45% in acc.@10, which confirms this metric is a much harder one but more reasonable to evaluate how well the synthesized sketch can resemble the corresponding image [14]. We can see that our full model can achieve comparative FG-SBIR results to the best model *Shortcut* (2.51% vs 2.98% in acc.@1 and 17.27% vs 17.43% in acc.@10). What's more, we can witness obvious contributions of each key component. Additionally, our model is also capable of a reverse task on sketch-to-image translation. Example qualitative results are shown in Fig. 4.

Fig. 4. Example sketch-to-image translation results.

V. CONCLUSION

In this paper, for the first time, we propose an unsupervised approach for image-to-sketch translation by developing a cycle+ consistency loss and exploiting the edge cue from real image, which resolve the problem of representation shifting caused by learning from unpaired image-sketch data. Experimental results on sketch recognition and SBIR illustrate the effectiveness of our method. In addition, our model is capable of sketch-to-image translation as well.

ACKNOWLEDGMENT

This work is supported by the National Natural Science Foundation of China (61601042, 61671078), 111 Project of China (B08004, B17007).

REFERENCES

[1] Y. Chen, S. Tu, Y. Yi, and L. Xu, "Sketch-pix2seq: a model to generate sketches of multiple categories," *CoRR*, 2017.
[2] M. Eitz, J. Hays, and M. Alexa, "How do humans sketch objects?" *SIGGRAPH*, 2012.
[3] A. Gonzalez-Garcia, J. van de Weijer, and Y. Bengio, "Image-to-image translation for cross-domain disentanglement," in *NeurIPS*, 2018.
[4] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio, "Generative adversarial nets," in *NeurIPS*, 2014.
[5] D. Ha and D. Eck, "A neural representation of sketch drawings," in *ICLR*, 2018.
[6] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz, "Multimodal unsupervised image-to-image translation," in *ECCV*, 2018.
[7] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *CVPR*, 2017.
[8] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *ECCV*, 2016.
[9] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *ICLR*, 2014.
[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NeurIPS*, 2012.
[11] H.-Y. Lee, H.-Y. Tseng, J.-B. Huang, M. Singh, and M.-H. Yang, "Diverse image-to-image translation via disentangled representations," in *ECCV*, 2018.
[12] M.-Y. Liu, T. Breuel, and J. Kautz, "Unsupervised image-to-image translation networks," in *NeurIPS*, 2017.
[13] Y. Qi, Y. Song, H. Zhang, and J. Liu, "Sketch-based image retrieval via siamese convolutional neural network," in *ICIP*, 2016.
[14] J. Song, K. Pang, Y.-Z. Song, T. Xiang, and T. M. Hospedales, "Learning to sketch with shortcut cycle consistency," in *CVPR*, 2018.
[15] J. Song, Q. Yu, Y. Song, T. Xiang, and T. M. Hospedales, "Deep spatial-semantic attention for fine-grained sketch-based image retrieval," in *ICCV*, 2017.
[16] X. Wang and A. Gupta, "Generative image modeling using style and structure adversarial networks," in *ECCV*, 2016.
[17] Q. Yu, F. Liu, Y. Song, T. Xiang, T. M. Hospedales, and C. C. Loy, "Sketch me that shoe," in *CVPR*, 2016.
[18] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *ICCV*, 2017.