Neurocomputing 322 (2018) 38–46

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Better freehand sketch synthesis for sketch-based image retrieval: Beyond image edges

Xianlin Zhang[a,*], Xueming Li[a,b], Xuewei Li[a], Mengling Shen[a]

[a] School of information and communication engineering, Beijing University of Posts and Telecommunications, China
[b] Beijing Key Laboratory of Network System and Network Culture, Beijing University of Posts and Telecommunications, China

ARTICLE INFO

Article history:
Received 12 February 2018
Revised 20 July 2018
Accepted 21 September 2018
Available online 26 September 2018

Communicated by Dr Jianjun Lei

Keywords:
Freehand sketch generation
CNN
Dual GAN
Improved faster R-CNN
Sketch recognition

ABSTRACT

With the rapid development of electronic touch screen and pressure-sensing devices, research on freehand sketches has become a hotspot in recent years. In this paper, we first propose a new freehand sketch generation model (FHS-GAN), which is based on the deep architecture of dual generative adversarial nets (GANs). We construct a model which utilize the deep convolutional neural network (CNN) and GAN to produce freehand sketches. We then propose an improved deep CNN model as a validated network, which is based on Faster R-CNN, to measure the similarity of real sketches and generated freehand sketches by FHS-GAN, and we test the improved model using the produced sketches with two large sketch datasets. The experiments show that the proposed FHS-GAN framework achieves state-of-the-art results in comparison with other baseline models. Furthermore, the generated sketches can be used for other sketch recognition tasks, such as in a pre-processing step for application in sketch-based image retrieval (SBIR) and fine-grained sketch-based image retrieval (FG-SBIR). Overall, our FHS-GAN model is important for the development of freehand sketches.

1. Introduction

Sketching comes naturally to humans, as it has been used for rendering the real world since early times [1]. With the proliferation of touch screens, we can now sketch effortlessly and ubiquitously by sweeping fingers on phones, iPads, smart watches, etc., and the study of hand-drawn sketches has thus become increasingly popular in recent years. However, how to automatically transform images into sketches, a problem that indeed has profound implications for sketch recognition tasks such as SBIR and FG-SBIR, still remains open.

"Sketch" is a wide concept in the computer vision (CV) field. For example, the exaggerated caricature sketches produced by artists (the caricature or cartoon sketches are exaggerated and have large deformations in shapes), the frontal human face sketches (this kind of sketch usually contains a detailed, rich facial texture structure) drawn by professionals, and freehand drawings by common users are all called "sketches". However, these three types of "sketches" are all very different in the richness of object details or the manner in which the sketches are produced (the first two types are normally drawn by professionals, and the third is generated in a casual way by anyone, even a child). The definition of a sketch in this paper refers primarily to the freehand sketch, which is produced by anyone with electronic touch-screen devices.

Plenty of prior work on sketches exist in the CV field, from the pioneering work of David Marr on primal sketches [2] to sketch classification [3–5]. Recent work on sketches includes freehand-drawn sketch synthesis [6–8], SBIR [9–11] and FG-SBIR [12,13]. Nonetheless, how to make machines draw sketches as humans do is still an open problem. Solving this problem opens the door for many sketch recognition tasks, e.g., generating sketches from real images can reduce the domain gap for SBIR and FG-SBIR.

This paper focuses on generating freehand sketches automatically from images by using models based on dual GAN learning. The contributions of the paper are mainly as follows:

(1) We develop a FHS-GAN model that can conduct sketch synthesis automatically in a simple way.
(2) We propose an improved CNN model based on Faster R-CNN as a similarity measure network for verifying the pseudo sketches generated by FHS-GAN, in addition, the improved structure can be used as a recognition model for sketch classification.
(3) The good results of the produced sketches can be used in both the field of recreation and for other sketch recognition tasks.

The rest of the paper is organized as follows: In Section 2, we provide the related work on sketch generation. In Section 3, we

* Corresponding author.
E-mail address: zxlin@bupt.edu.cn (X. Zhang).

https://doi.org/10.1016/j.neucom.2018.09.047
0925-2312/

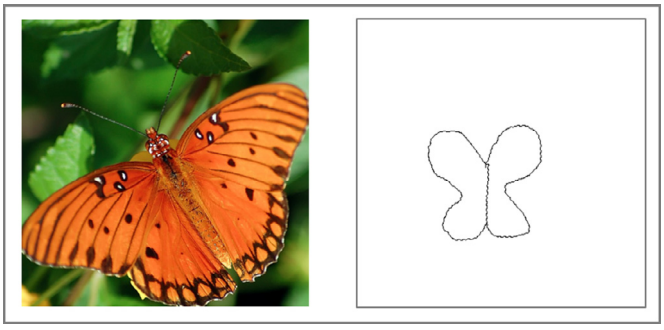


Fig. 1. An example of a natural butterfly image and the freehand sketch.

provide a detailed review of the proposed FHS-GAN model and the validated network. In Section 4, we describe the experiment settings and then analyze the sketch generation results. Conclusions on the freehand sketch generation model proposed in the paper are presented in Section 5.

2. Related work

Sketch generation, which is intended to process a real image, produces a sketch image similar to a freehand one (an example of a real image and a freehand sketch is shown in Fig. 1). Generating sketches from images has usually been done for two main sketch applications. One is as a pre-processing step for other sketch-related tasks (e.g., SBIR and FG-SBIR) to bridge the domain gap, and the second is just for recreation. In addition, most of the sketch generation algorithms are proposed for the first application. Therefore, methods of sketch generation have some overlap with object detection and edge extraction tasks. Regarding the prior sketch generation algorithms, they can mainly be divided into two types: generation methods based on manual features and models of deep architecture features.

2.1. Prior work

2.1.1. Work based on manual features

Most of the studies on automatic sketching employed a contour detection and object segmentation approach [14–17], which is aimed to produce curves that perfectly depict an image or constitute global object profiles. Zhu et al. [14] exploited the inherent topological structure of salient contours. Wang et al. [15] introduced a morphing technique for sketches by using skeleton and order curves. Arbelaez et al. [16] proposed a general framework to transform the output of any contour detector into a hierarchical region tree with the intent to generate object contours that are most similar to human image segmentation. Marvaniya et al. [17] proposed to sketch an object given a set of images of the same category. The essential idea was to discover repeatable salient contours across a set of images of the same class. Later, Lim et al. [7] proposed a novel approach to both learning and detecting local contour-based representations for mid-level features named sketch tokens. The sketch tokens were learned using supervised mid-level information in the form of hand-drawn contours in images. In contrast, Qi et al. [18] proposed a method exploiting perceptual grouping principles to produce a sketch from a single image, which made sketch generation more convenient and applicable. Cheng et al. [19] proposed a real-time system of hierarchical feature selection (HFS) to perform image segmentation utilizing different feature setting in different scale levels. The HFS system can achieve a comparative segmentation result with GPU speed up.

2.1.2. Work based on deep features

The methods mentioned above mainly utilize hand-crafted features from images. However, with the rapid development of the CNN and fully convolutional network (FCN) frameworks, the theory of deep learning (DL) has achieved great success in the CV field. Liu et al. [20] constructed a novel method of deep embedding learning (DEL) that can transform superpixels into image segmentation more efficiently compared to HFS and it was the first work to perform image segmentation using an end-to-end deep learning system. In addition, Xie and Tu [21] developed a new holistically nested edge detection (HED) model that addressed two important issues: (1) holistic image training and prediction and (2) multi-scale and multi-level feature learning. Their model performed image-to-image prediction by means of a deep learning model that leveraged the FCN model and deeply supervised nets. Since CNNs have been proved to be effective for this task, Liu et al. [22] proposed an accurate richer convolution features (RCF) model based on HED. The proposed network fully exploited multi-scale and multilevel information on objects to perform image-to-image prediction by combining all the meaningful convolutional features in a holistic manner.

However, Yu et al. [23] designed an end-to-end deep architecture, which was named CASENet, for category-aware semantic edge detection. Later, Liu et al. [24] proposed a novel FCN architecture using diverse deep supervision (DDS) within a multi-task framework for semantic edge detection. The lower layers of the model were responsible for the generation of category-agnostic edges, while the high layers aimed at detecting of category-aware semantic edges.

2.2. Our work

Most of the sketch generation methods are based on the image domain itself, which means that sketches are synthesized by using information contained in the original images without considering the feature from the sketch domain. However, freehand sketches usually consist of only a few simple strokes and have very few detail features. Therefore, it is simply easier to use the characteristics of an image's shape, contour, or texture to produce sketches similar to those drawn by humans. One study [25] has proved that the object boundaries are not satisfactory for freehand sketches. The view that human-drawn sketches are very different from object edges or boundaries can indeed be seen from Fig. 1 as well. Therefore, we should consider using the content of freehand sketches to produce pseudo sketches from images. With the development of GANs, which was proposed by Goodfellow et al. [26] in 2014, the GAN family has made much progress in the past two years in cross-domain image-to-image translations, especially those that utilized FCNs and conditional GANs (cGANs) [27] to enable a unified treatment of these tasks. We then naturally propose the idea of developing an unsupervised learning framework for general-purpose image-to-freehand sketch translation.

Our approach is inspired by dual GAN learning [28] and Wassertein GAN learning [29]. Dual learning means training two "opposite" translators simultaneously by minimizing the reconstruction loss resulting from a nested application of the two translators. Our FHS-GAN model develops an automatic framework for image-to-freehand sketch translation, and the generated sketches can be used in both the applications of recreation and the procedure of pre-processing steps for SBIR and FG-SBIR tasks. The proposed model differs from the original dual GAN of Xia [28] in two main aspects. First, the process of training is hard, as freehand sketches and images vary greatly, so we obtain a set of parameters by modifying the loss function, which can produce adequate sketches at last. Second, we develop an improved CNN model based on Faster R-CNN [30] to distinguish and evaluate the

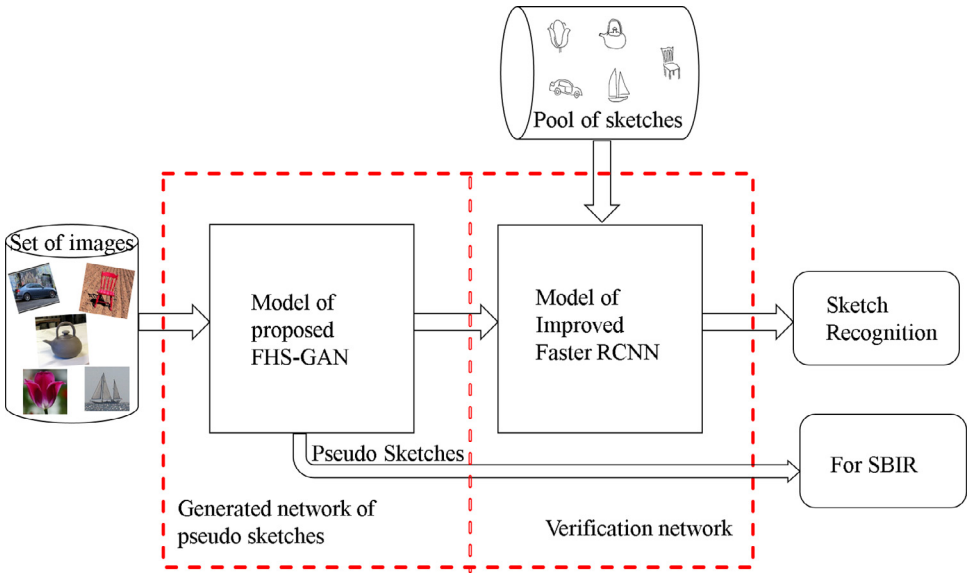


Fig. 2. The framework of the proposed work.

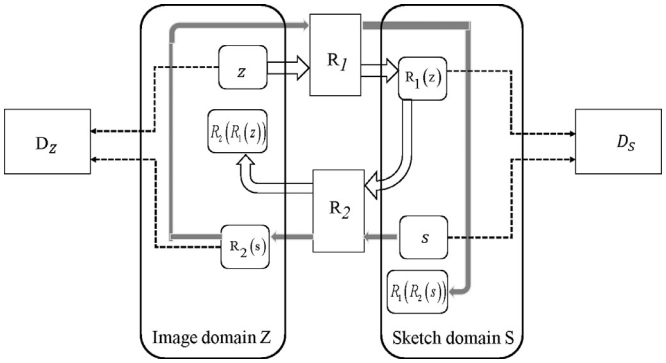


Fig. 3. The framework of the sketch generation model.

quality of real and generated sketches. The whole framework of our architecture is shown in Fig. 2, the consist part of FHS-GAN and improved Faster R-CNN are described in detail next.

3. Proposed method

3.1. Method of the FHS-GAN network

The motivation of the FHS-GAN network originates from the idea of GAN and its derivative algorithms, such as CGAN [31], DCGAN [32], infoGAN [33], cycle GAN [34], WGAN [29], and dual GAN [28].

The basic concepts and theories about GANs can be found in the literature on GAN [26]. The work on FHS-GAN in this paper is similar to Refs. [34] and [28], as the original GAN had the problem of difficulty in training the network and was a semi-supervised mechanism model, which was not very convenient. However, the work in [34] and [28] was performed in an unsupervised way, and the central idea is to focus on obtaining two mapping functions in two different media-data domains, which means training two sets of models (two generators and discriminators) at the same time. Thus, training a bi-directional network sounds more suitable for our freehand sketch generation task. The architecture of FHS-GAN is shown in Fig. 3 and details are presented below.

Formulation As shown in Fig. 3, $\{z_k\}_{k=1}^{K}$ and $\{s_j\}_{j=1}^{J}$ are given training examples, which are from two varying domains of images Z and freehand sketches S ($s_j \epsilon S$, $z_k \epsilon Z$). Traditionally, a dual model involves learning two mapping functions between two domains, that is, $R_1: Z \rightarrow S$ and $R_2: S \rightarrow Z$. The goal of the model is to train two optimized mapping $R_1(z)$ and $R_2(s)$ to generate outputs to "fool" the corresponding discriminators D_S and D_Z. For discriminator D_S, it discriminates between real freehand sketch s and the generated one $R_1(z)$ during the training process; Discriminator D_Z departs z and $R_2(s)$ in the same way. At the end of training, we aim to obtain two optimized discriminators.

Objective In the proposed FHS-GAN model, the loss terms consist of two parts in total. One is the adversarial loss, and the other is the reconstruction loss. Take our task into consideration, we have sketch and image samples in two different domains, and we would like to obtain better generated freehand sketches from the side of $z \rightarrow R_1(z)$, but not the opposite side result of the dual work model. Different from the loss of dual GAN [28]: (1) We focus on training one side of the dual model, which makes the produced sketches from generator $R_1(z)$ better, (2) We strengthen the constraints of objective function utilizing the term of gradient penalty, as deep structures based on GANs often have the problem of harding training and slow convergence. Therefore, in the FHS-GAN model, the objective loss is mainly defined as

$$L_1 = L_{GAN}(R_1, D_S, Z, S) \tag{1}$$

$$L_2 = L_{GAN}(R_2, D_Z, S, Z) \tag{2}$$

$$L_{GAN}(R_1, D_S, Z, S) = E_{s\sim p_d(s)}[D_S(s)] - E_{z\sim p_d(z)}[D_S(R_1(z))] + \delta E_{\hat{z}\sim p(\hat{z})}\left[\left(\left\|\nabla D_S(\hat{z})\right\|_2 - 1\right)^2\right] \tag{3}$$

$$L_{GAN}(R_2, D_Z, S, Z) = E_{z\sim p_d(z)}[D_Z(z)] - E_{s\sim p_d(s)}[D_Z(R_2(s))] + \delta E_{\hat{s}\sim p(\hat{s})}\left[\left(\left\|\nabla D_Z(\hat{s})\right\|_2 - 1\right)^2\right] \tag{4}$$

$$L_C = \beta \|z - R_2(R_1(z))\| \tag{5}$$

$$L(R_1, R_2, D_Z, D_S) = L1 + L2 + L_C \tag{6}$$

$$R_1^*, R_2^* = \arg\min_{R_1,R_2}\max_{D_Z,D_S} L(R_1, R_2, D_Z, D_S) \tag{7}$$

Eqs. (1)and (2) are the adversarial losses. L_1 expresses the loss from the domain Z to S and L_2 means the loss from the opposite mapping S to Z. Specific details of L_1 and L_2 are expressed by formula (3) and (4). The last term in Eqs. (3) and (4) is the added gradient penalty term, which is different to dual GAN, for random samples $\hat{z} \sim p(\hat{z})$ (the same to $\hat{s} \sim p(\hat{s})$) to improve the training stability. $p(\hat{z})$ means sampling uniformly along straight lines between pairs of points sampled from the data distribution $p_{d(z)}$ and the generator distribution $p_{d(s)}$ in Eq. (3). $p(\hat{s})$ is the same meaning in formula (4), δ is a hyper-parameter. Further, the loss formats in Eqs. (3) and (4) that we used are the form of Wasserstein distance [29] rather than the log form of the original GAN [26]. Eq. (5) presents the recovery loss. For this loss, we just consider the side of generating sketches, which means emphasizing the one way of $z \rightarrow R_1(z)$ for the construction loss and β is a balance parameter. Formula (6) exhibited the overall form of the objective function.

In the preliminary experiments, we use L_1 and L_2 distances to train the FHS-GAN network respectively, and the experimental results show that there is no obvious difference between the two distances in improving the final generation performance. We choose L_1 distance in the network setting and parameters δ in (1) and β in (3) are set to 10 and 20, respectively.

FHS-GAN Network and Training For the generator of FHS-GAN, we employ a six-layer convolutional neural network to produce sketches as our task is an end-to-end training and the output is a 2D sketch. In order to get better data features using the information of training data, we choose to configure the network layers into a U-shape net such as DenseNet [35] in the generator model. Such a design enables different level information to be shared adequately between input and output, which is beneficial to improve the generation results since our image-to-sketch translation problem implicitly assumes alignment between data structure in the input and output. The configuration of the two discriminators are the same on our dual model, we use a four layer CNN to construct the network which is similar in [36]. We run the discriminator convolutionally across the image, averaging all responses to provide the ultimate generation. The architecture of discriminator is proven to require fewer parameters, and has no constraints over the dimension of the input image. During our training procedure, the size of the input and output of the training data is 256×256.

During the training of FHS-GAN model, we choose to use the optimization algorithm of RMSProp, which can be found in Ref. [37], to prevent the vanishing of gradients and to better optimize the dual network. In addition, we utilize the deep framework of Tensorflow on a server computer equipped with 4.2 GHz and 32G CPU. The hardware of the experiments is GeForce GTX TITAN X with Cudnn speedup.

3.2. Method of improved Faster R-CNN network

We have introduced the FHS-GAN model which can produce pseudo hand-drawn sketches in the above section, and it naturally comes the question that what about the quality of the generated ones compared to real hand-drawn sketches? It is also one of the major problems that have not been explicitly resolved in most image-to-image translation works. As the specific sketch application of our task to solve this problem, we propose to construct a deep sketch recognition network as a verification and similarity measure model for the first time.

Considering Faster R-CNN has been proven to be useful and powerful capacity in object localization and detection in recent years, we aim to construct a deep model for both recognize a variety of hand-drawn sketches and perform qualitative verification experiments on the pseudo sketches generated by the proposed FHS-GAN model. In this paper, we construct an improved architecture based on Faster R-CNN with ZF model [38]. The whole framework of the improved network is shown in Fig. 4.

Original network : In order to facilitate the following introduction of the improved network, we give a brief presentation on Faster R-CNN. As shown in Fig. 4, Faster R-CNN (with ZF) model has three parts, image features are first extracted by ZF network, then the generated feature maps are sent to the region proposal network (RPN) to obtain the region proposals, and they go into the Fast R-CNN [39] net to finally get the bounding box and classification prediction score respectively. The ZF net in the first part is a deep network with 8 layers (and has 5 convolution layers) to obtain getting deep features of input data, and the RPN net in the second part is a one-convolution layer network for the selection of candidate boxes. It can also be regarded as a classification net for producing proposals. More detailed introduction can be seen in the literature of Fast R-CNN [39] and Faster R-CNN [30].

It is known from the above introduction that feature maps extracted by ZF net are sent to a one-layer convolution RPN net for prod ucing candidate proposals, the results of RPN are then sent into the classification layer (Cls layer) and the positioning layer (Bbox layer).

Improved network : One of the most important problems affecting the location and recognition results of Faster R-CNN is the accuracy of regional proposals which generated from the RPN network. However, RPN has only one convolution layer in the Faster R-CNN structure which can not obtain multi-scale features from the previous step. Different to the original architecture, the inception module of GoogleNet [40] has two obvious advantages: one is to allow a significant increase in the number of units per step, and the computational complexity will be controlled as the general usage of dimension reduction can protect the large number of input filters from the last step to the next layer and reduce the dimensions before they are deal with a large patch size. The other is that visual information is processed on different scales and then aggregated, so that the next step can extract features from different scales at the same time. In addition, the speed is faster using the inception module. In order to make full use of the multi-scale features of the image to improve the prediction performance of region proposals, we propose to extend the RPN network into a new parallel architecture inspired by GoogleNet [40].

The modified structure which we named RPN-X is raised to replace original RPN. As presented in Fig. 4, the new network architecture consists of four parallel convolution layers, and convolution layers $2-1$, $3-1$, and $4-1$ are respectively connected to the convolution layers 2, 3, and 4 afterwards. The specific parameters of the network are : The kernel size of four parallel convolution (conv1, conv2, conv3 and conv4) layers is 3×3, 1×1, 1×1 and 1×1. For layers of $2-1$, $3-1$, and $4-1$, kernels of these three convolution layers are 3×3, 5×5, and 1×1.

We train a sketch recognition model based on this improved Faster R-CNN network. The experimental results present an approximate of 4–6 percentage improvements compared to the original network in freehand sketch recognition on the largest sketchy database [25] which show the effectiveness of parallelism RPN network in predicting bounding box and recognizing freehand sketches. We give a detailed description of the recognition accuracy on sketches in the upcoming analysis of the experiment results.

4. Experiments

In this part, we will demonstrate the generated results from two aspects. First is the intuitive experimental sketch results generated by proposed FHS-GAN model, and second is the verified experimental results and analysis. In addition, in order to verify

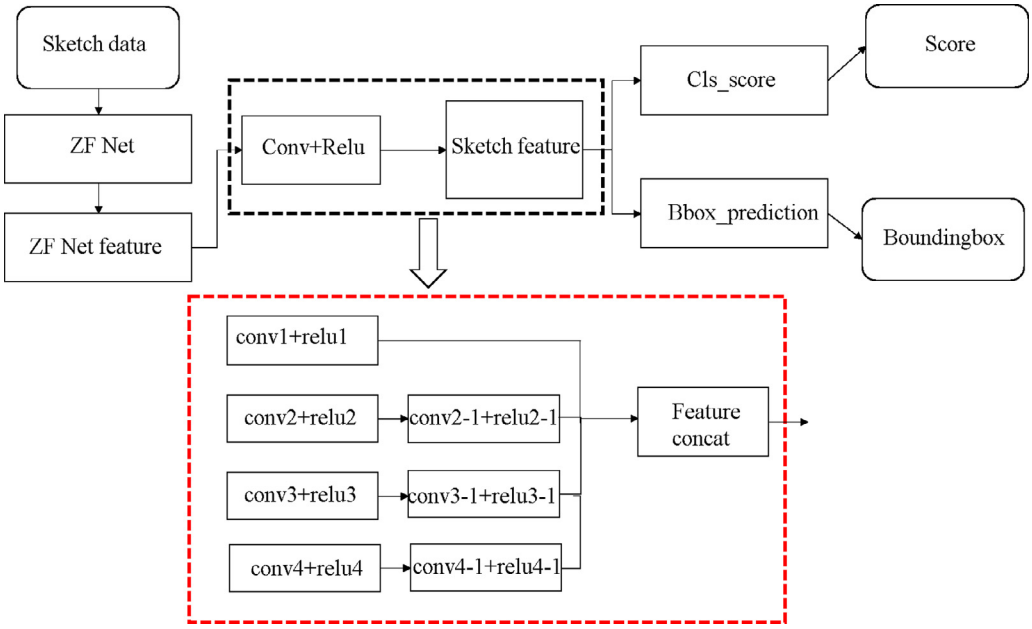


Fig. 4. The framework of improved Faster R-CNN, the black dotted box is the original network, and the red dotted box is the improved module. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

the quality of the sketch and the good adaption of the model, we identify the generated sketches from two sketch databases.

4.1. Intuitive results

We prepare unpaired data from the domain of sketches and images, which are all from the Sketchy database [25] to train the FHS-GAN model, and Sketchy [25] is the largest sketch database so far, with a total number of about 70,000 (The database contains a total of 125 classes, and 100 images for each class). The training images and sketches in our experiments are 10,000 from 50 different categories (both the number of sketches and the number of images are 5000, and the categories include standing bird, shoe, airplane, ape, starfish, etc.), and there are 500 test images. The initial learning rate is 0.002, and the decay ratio and the epochs are set to 0.9 and 20, respectively. Finally, the FHS-GAN network is used to obtain the model that can automatically generate sketches. We compared the generated sketches of the original DualGAN and our proposed FHS-GAN model during the experiments. Part of the training sketches and intuitive generated sketches are exhibited in Fig. 5.

From the intuitive results in Fig. 5, the sketches of DualGAN are not stable, and appear incomplete contours and missing lines as the generated bird and shoes. However, sketches using FHS-GAN are generated with high quality and very similar both in the smoothness of sketch strokes and the overall shape of real sketches. The generated sketches are robust despite the cluttered background and it is obvious that they are better than the results without considering the characters of sketches. Furthermore, the generated pseudo sketches are fine-grained, such as the butterfly exhibited in Fig. 5, which can be well used for other recognition tasks as SBIR and FG-SBIR for reducing the domain gap. The quantitative results show that FHS-GAN model can produce high-quality sketches and is effective.

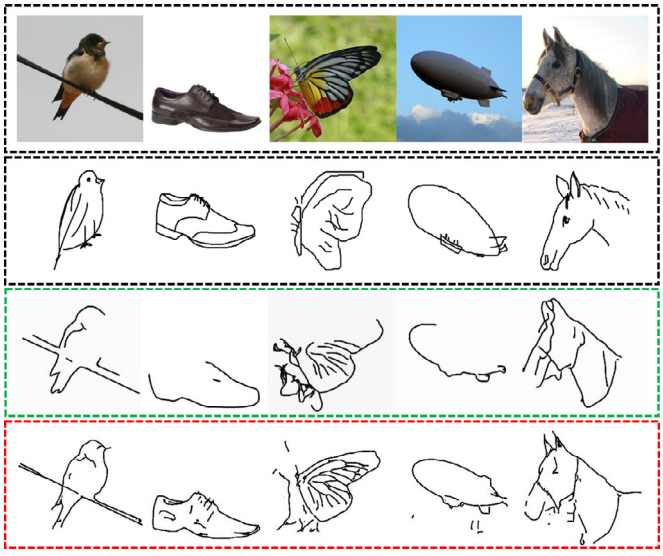


Fig. 5. Examples of sketches generated by our FHS-GAN model. The first row is the original images from the Sketchy database, and the categories are standing bird, shoe, butterfly, blimp and horse. The second row is the freehand sketches corresponding to the images in Sketchy. The third row, which are in green dash rectangle, is the sketches generated using DualGAN, and the fourth row is the generated results by our proposed FHS-GAN. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

4.2. Verification experiments

4.2.1. Recognition on the Sketchy database

The whole idea of verification is that we first develop two freehand sketch deep models using Faster R-CNN and our improved Faster R-CNN for recognition of the sketches, and then, we verify the generated sketches by using the trained deep sketch model to prove recognition performance.

Therefore, the first step is to obtain a deep sketch recognition model (we take the improved model as an example to illustrate the process) using the improved Faster R-CNN with ZF net. We choose to utilize about 40,000 varied sketches from 50 categories of Sketchy for the training of our deep net. We then obtain a deep recognition model to classify freehand sketches and achieve good performance in terms of the classification accuracy and average precision (AP). We list the AP values for all 50 classes and the

Table 1
Parts of classification accuracy values and the overall AP of two models ('ori' means the Faster R-CNN model, 'improve' presents the improved Faster R-CNN).

Image category	Precision (ori)	Precision(improve)	AP (ori)	AP(improve)
Bicycle	0.930	0.963	––	––
Hot-air balloon	0.924	0.957	––	––
Sailboat	0.902	0.935	––	––
Cabin	0.897	0.935	––	––
Chair	0.893	0.933	––	––
Butterfly	0.573	0.664	––	––
Tiger	0.554	0.613	––	––
Bear	0.417	0.476	––	––
Lion	0.392	0.435	––	––
Dog	0.297	0.321	––	––
––	––	––	**0.7859**	**0.8363**

Table 2
Parts of classification accuracy values and the overall AP on generated sketches by FHS-GAN and real sketches.

Image category	Cla-precision(generated)	Cla-precision(real)
Sailboat	0.935	0.93
Hot-air balloon	0.74	0.95
Helicopter	0.67	0.87
Blimp	0.58	0.80
Rifle	0.57	0.86
tree	0.56	0.72
Zebra	0.53	0.76
Cup	0.48	0.83
Flower	0.48	0.7
Pear	0.46	0.87
AP	**0.60**	**0.83**

classification accuracy for a few sketch categories for which the values are the top and bottom 5. In order to verify the efficiency of the improved Faster R-CNN, we list the classification precision of the original Faster R-CNN model as well. The results are shown in Table 1.

From Table 1, it is obvious to show that the improved model is effective in terms of classification accuracy. The AP is about 5 percentage higher on the overall 50 categories compared to Faster R-CNN, which has proven the added module is useful. In addition, the recognition result for different classes of Sketchy varies considerably. This may be due to: (1) The huge differences in the abstraction of the target, the drawing skill, etc. when people made the sketches. (2) The constructed model still has limitations in perfectly detecting and recognizing sketches. However, we have reason to believe that the trained deep model can recognize freehand sketches well overall as the AP is up to 0.836.

4.2.2. Recognition of generated sketches

To measure the similarity between the generated sketch and the real sketch, and to verify the domain adaption of the constructed FHS-GAN model, we conduct a two-part experiments with two large sketch databases. (1) Using the constructed similarity measure network (the sketch recognition model based on improved Faster-RCNN) to classify the generated pseudo sketches of ten categories randomly sampled from Sketchy database for recognition verification. (2) We choose ten overlapping image

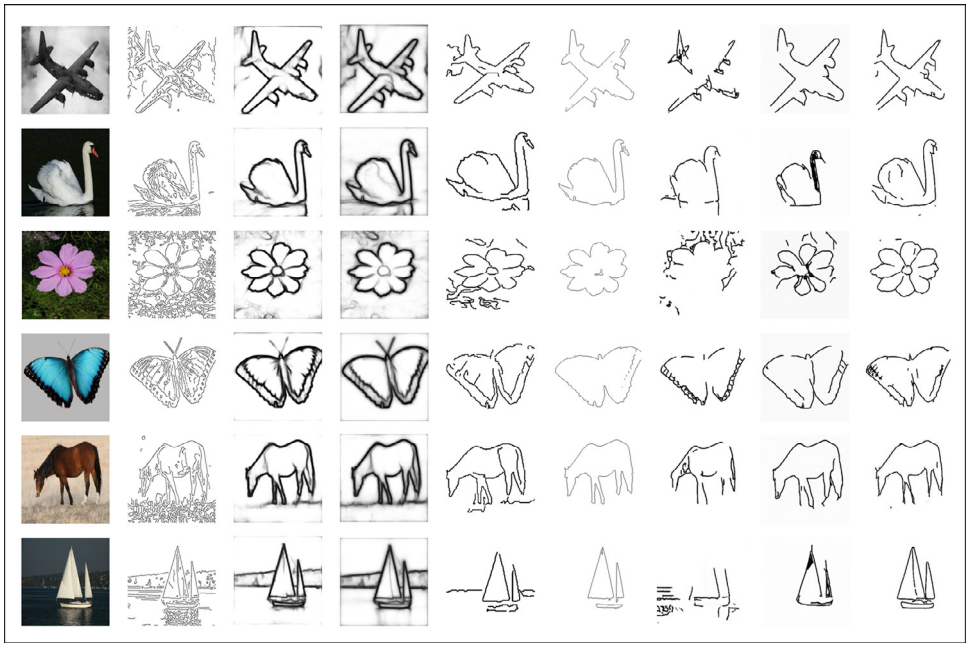
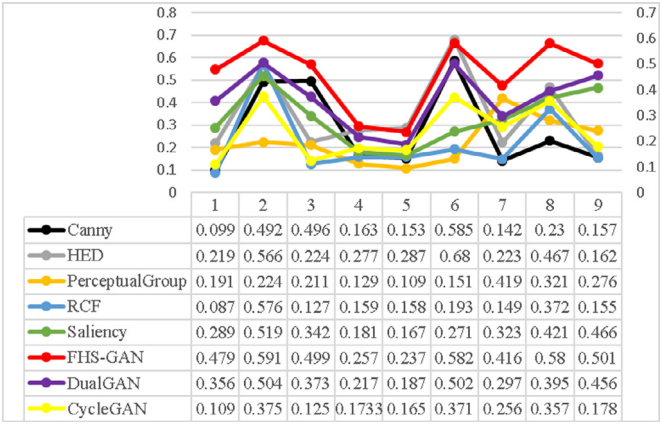


Fig. 6. Examples of generated 'sketches'. The first column (a) shows the original images from Flickr15k, and the image categories are airplane, swan, flower, butterfly, horse and sailboat. The second column (b) is the results from Canny. The 'sketches' in the third column (c) are from the model of HED. The results in the fourth column (d) are from RCF, (e) are the results of perceptual grouping. The fifth column (f) is based on the method of saliency, column (g) presents the pseudo sketches by CycleGAN, the eighth column (h) is the results using DualGAN and the last column is the sketches generated by the proposed FHS-GAN model.



	1	2	3	4	5	6	7	8	9
Canny	0.099	0.492	0.496	0.163	0.153	0.585	0.142	0.23	0.157
HED	0.219	0.566	0.224	0.277	0.287	0.68	0.223	0.467	0.162
PerceptualGroup	0.191	0.224	0.211	0.129	0.109	0.151	0.419	0.321	0.276
RCF	0.087	0.576	0.127	0.159	0.158	0.193	0.149	0.372	0.155
Saliency	0.289	0.519	0.342	0.181	0.167	0.271	0.323	0.421	0.466
FHS-GAN	0.479	0.591	0.499	0.257	0.237	0.582	0.416	0.58	0.501
DualGAN	0.356	0.504	0.373	0.217	0.187	0.502	0.297	0.395	0.456
CycleGAN	0.109	0.375	0.125	0.1733	0.165	0.371	0.256	0.357	0.178

Fig. 7. The recognition precision value of each category. Numbers 1–9 on the *x*-axis represent the image categories of horse, hot-air balloon, flower, butterfly, bicycle, airplane, sailboat, swan and duck, respectively. The table shows the recognition results of different models for each category.

categories with Sketchy and Flickr15k [10], and then test the domain adaptation on Flickr15k database using the proposed FHS-GAN model.

(1). We randomly select ten image categories in the Sketchy database for classification, and also use the classification precision (cla-precision) and the average accuracy to represent the recognition results. The experimental results are shown in Table 2.

From Table 2, we can see that the classification results of the recognition model on the pseudo sketches of ten random categories are worse than those of the corresponding real freehand sketches. This may be due to: (a). The real sketches vary greatly, and they are lack of useful information. Real images can not learn well from sketches, thus affecting the recognition efficiency. (b). There may be multiple targets with clutter backgrounds in the real image, but the corresponding sketch is usually a single and clean target. Therefore, the pseudo sketch generated from the image inevitably produces irrelevant noise affecting the recognition results. However, the recognition accuracy is still competitive based on the AP value. It proves that the proposed sketch generation model can generate better sketch results.

(2). In the process of SBIR and FG-SBIR, utilizing the edges acquired from original images as a pre-processing step to reduce the domain gap is common. However, freehand sketches are not image edges, as sketches are casual and simple, with no rules at all. This point of view is also mentioned in Ref. [25].

As we have obtained the deep model to recognize freehand sketches, we then compare the sketches obtained by FHS-GAN with several other algorithms, including models for producing both pseudo sketches and image edges by using the deep recognition model to verify that our results are superior.

To ensure generality and fairness to all algorithms, we perform the validation experiments on the same nine categories of images both in the Sketchy database and the Flickr15k [10] dataset. For more details, the baselines we used for validation include the methods of Canny [41], perceptual grouping [42], saliency based on a random model [43], HED [21], RCF [22], DualGAN [28], and the CycleGAN [34]. Images of the nine classes include airplane, horse, flower, butterfly, sailboat, swan, bicycle, hot-air balloon and duck. The total number of testing images is 1206. Parts of the generated 'sketches' by different models are presented in Fig. 6.

Canny is a classic operator to extract the image edges, but the generated results are noisy and multi-object that can be seen in Fig. 6. The results are similar using HED and RCF at some extent as the model of RCF is improved based on HED. They can produce relatively complete image edges but lack of sketch-style as they are obtained by using the information of images only. Pseudo sketches generated by perceptual grouping are slightly like freehand sketches. But some results still contain redundant information, such as the generated flower. Sketches created by saliency are simple and clean, but they have no fine-grained feature of the object. However, results generated using adversarial networks are more sketch-style since we use real sketches as training data

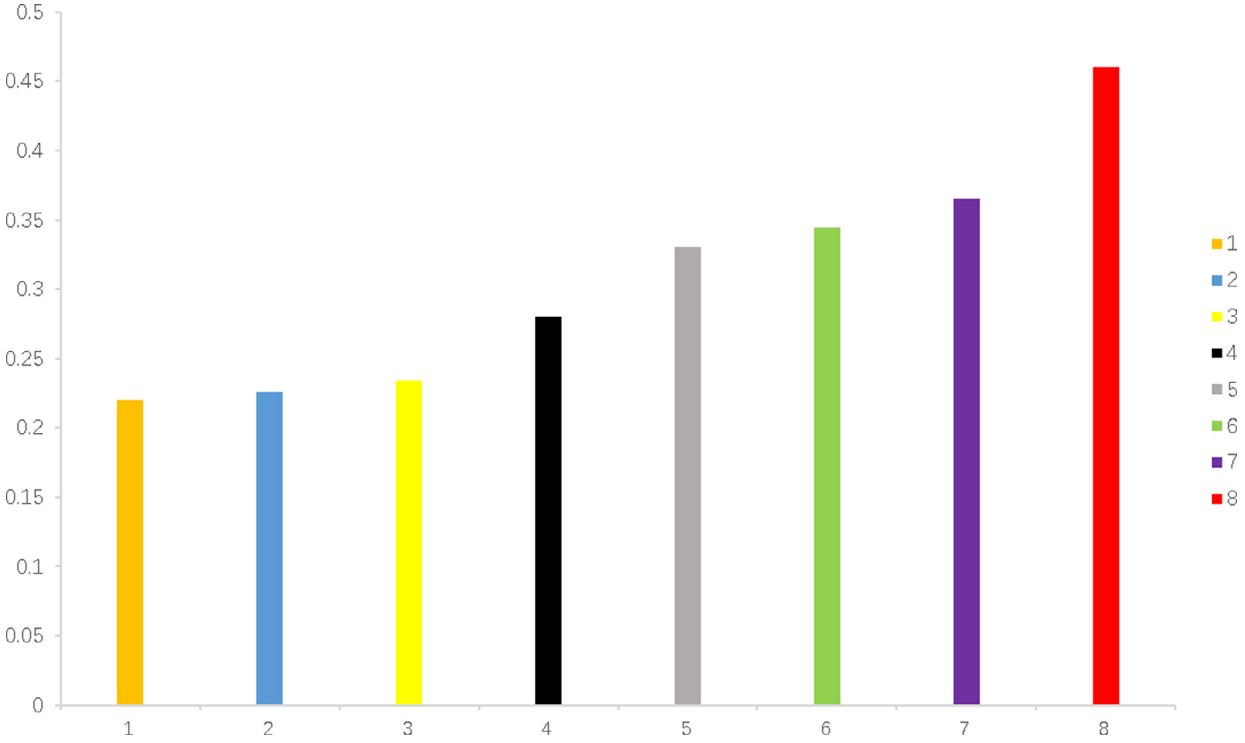


Fig. 8. The overall AP value of the nine categories. Numbers 1–8 on the *x*-axis represent the algorithms of perceptual grouping, RCF, Canny, HED, saliency, CycleGAN, DualGAN and the proposed FHS-GAN, respectively. Numbers 0–0.5 on the *y*-axis present the value of AP.

that are unlike methods based on the edges or outlines of images. Sketches generated by FHS-GAN are more stable and complete, which can be obviously seen in Fig. 6, compared with DualGAN and CycleGAN. The intuitive results reveal that the proposed model is effective in generating sketches.

Figs. 7 and 8 show the classification accuracy value of each category and the AP of all 9 categories.

From Fig. 7, it is obvious that different algorithms have good recognition of certain categories. For example, the HED model is efficient in the airplane category, the RCF model performs better in the hot-air balloon category, etc. The reason why the RCF model did not perform well may be due to the fact that the predicted bounding boxes could not locate the object in the whole sketch region or lack of preprocessing. However, our method performs well on all nine categories. The models are performed relatively better by using GANs as a whole, which is shown in Fig. 8. Furthermore, the best AP value achieved for all nine categories is 0.46 (note: no preprocessing steps are performed on all generated sketches) for the cross dataset by our proposed FHS-GAN model.

5. Conclusion

Freehand sketches are becoming more important with the proliferation of digital devices. Therefore, generating sketches automatically is an urgent need for many applications. We propose a FHS-GAN, which is based on the dual GAN learning, to produce sketches. In addition, we train a better deep sketch model, which is based on Faster R-CNN, and use the largest sketch database at present to recognize freehand sketches. The verification experiments between the proposed method and the baseline models prove that our FHS-GAN can yield state-of-the-art results. In addition, the produced sketches can be well used for other sketch-based tasks as SBIR and FG-SBIR.

Acknowledgment

We gratefully acknowledge the support of the NVIDIA Corporation with the donation of the TIAN X GPU used for our deep learning.

References

[1] M. Eitz, J. Hays, M. Alexa, How do humans sketch objects? ACM Trans. Gr. 31 (4) (2012) 1–10, doi:10.1145/2185520.2335395. http://dl.acm.org/citation.cfm?doid=2185520.2335395.

[2] D. Marr, E. Hildreth, Theory of edge detection, Proc. R. Soc. B Biol. Sci. 207 (1167) (1980) 187–217, doi:10.1098/rspb.1980.0020. http://rspb.royalsocietypublishing.org/cgi/doi/10.1098/rspb.1980.0020.

[3] Q. Yu, Y. Yang, Y.-Z. Song, T. Xiang, T. Hospedales, Sketch-a-Net that Beats humans, in: Proceedings of the British Machine Vision Conference (BMVC)(ii) (2015) 1–11. doi:10.5244/C.29.7.http://arxiv.org/abs/1501.07873.

[4] Q. Yu, Y. Yang, F. Liu, Y.Z. Song, T. Xiang, T.M. Hospedales, Sketch-a-Net: a deep neural network that beats humans, Int. J. Comput. Vis. 1 (i) (2016) 1–15, doi:10.1007/s11263-016-0932-3.

[5] R.G. Schneider, T. Tuytelaars, Sketch classification and classification-driven analysis using Fisher vectors, ACM Trans. Gr. 33 (6) (2014) 2691–2698, doi:10.1145/2661229.2661231. http://dl.acm.org/citation.cfm?doid=2661229.2661231.

[6] Y. Qi, J. Guo, Y. Li, H. Zhang, T. Xiang, Y.Z. Song, Sketching by perceptual grouping, in: Proceedings of the IEEE International Conference on Image Processing, ICIP 2013, 2013, pp. 270–274, doi:10.1109/ICIP.2013.6738056.

[7] J.J. Lim, C.L. Zitnick, P. Dollar, Sketch tokens: a learned mid-level representation for contour and object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, in: CVPR '13, IEEE Computer Society, Washington, DC, USA, 2013, pp. 3158–3165, doi:10.1109/CVPR.2013.406.

[8] Y. Li, T.M. Hospedales, Y.-Z. Song, S. Gong, Free-hand sketch recognition by multi-kernel feature learning, Computer Vision and Image Understanding 137 (1077-3142) (2015) 1–11, doi:10.1016/j.cviu.2015.02.003. http://www.sciencedirect.com/science/article/pii/S1077314215000375.

[9] Y. Qi, Y.Z. Song, H. Zhang, J. Liu, Sketch-based image retrieval via Siamese convolutional neural network, in: Proceedings of the International Conference on Image Processing, ICIP, 2016, pp. 2460–2464, doi:10.1109/ICIP.2016.7532801.

[10] R. Hu, J. Collornosse, A performance evaluation of gradient field HOG descriptor for sketch based image retrieval, Comput. Vis. Image Underst. 117 (7) (2013) 790–806, doi:10.1016/j.cviu.2013.02.005.

[11] M. Eitz, K. Hildebrand, T. Boubekeur, M. Alexa, Sketch-based image retrieval: benchmark and bag-of-features descriptors, IEEE Trans. Visual. Comput. Gr. 17 (11) (2011) 1624–1636, doi:10.1109/TVCG.2010.266.

[12] Q. Yu, F. Liu, Y. Song, T. Xiang, T.M. Hospedales, C.C. Loy, Sketch me that shoe, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 00, 2016, pp. 799–807, doi:10.1109/CVPR.2016.93.

[13] K. Li, K. Pang, Y.-Z. Song, T.M. Hospedales, T. Xiang, H. Zhang, Synergistic Instance-level subspace alignment for fine-grained sketch-based image retrieval, IEEE Trans. Image Process. 26 (12) (2017) 5908–5921, doi:10.1109/TIP.2017.2745106. http://ieeexplore.ieee.org/document/8016664/.

[14] Q. Zhu, G. Song, J. Shi, Untangling cycles for contour grouping, in: Proceedings of the Eleventh IEEE International Conference on Computer Vision, IEEE, 2007, pp. 1–8, doi:10.1109/ICCV.2007.4408929. http://ieeexplore.ieee.org/document/4408929/.

[15] S. Wang, T. Kubota, J.M. Siskind, J. Wang, Salient closed boundary extraction with ratio contour, IEEE Trans. Pattern Anal. Mach. Intell. 27 (4) (2005) 546–561, doi:10.1109/TPAMI.2005.84.

[16] P. Arbelaez, M. Maire, C.C. Fowlkes, J. Malik, Contour detection and hierarchical image segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 33 (5) (2011) 898–916.

[17] S. Marvaniya, S. Bhattacharjee, V. Manickavasagam, A. Mittal, Drawing an Automatic Sketch of Deformable Objects Using Only a Few Images, in: A. Fusiello, V. Murino, R. Cucchiara (Eds.), Computer Vision – ECCV 2012. Workshops and Demonstrations, Springer Berlin Heidelberg, 2012, pp. 63–72, doi:10.1007/978-3-642-33863-2_7. http://link.springer.com/10.1007/978-3-642-33863-2_7.

[18] Y. Qi, Y.Z. Song, T. Xiang, H. Zhang, T. Hospedales, Y. Li, J. Guo, Making better use of edges via perceptual grouping, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1856–1865, doi:10.1109/CVPR.2015.7298795.

[19] M.-M. Cheng, Y. Liu, Q. Hou, J. Bian, P. Torr, S.-M. Hu, Z. Tu, HFS: hierarchical feature selection for efficient image segmentation, in: B. Leibe, J. Matas, N. Sebe, M. Welling (Eds.), Proceedings of the European Conference on Computer Vision – ECCV 2016, Springer International Publishing, Cham, 2016, pp. 867–882.

[20] Y. Liu, P.-T. Jiang, V. Petrosyan, S.-J. Li, J. Bian, L. Zhang, M.-M. Cheng, Del: deep embedding learning for efficient image segmentation, in: Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI), 2018.

[21] S. Xie, Z. Tu, Holistically-nested edge detection, Int. J. Comput. Vis. 125 (1–3) (2017) 3–18, doi:10.1007/s11263-017-1004-z. http://link.springer.com/10.1007/s11263-017-1004-z.

[22] Y. Liu, M.-M. Cheng, X. Hu, K. Wang, X. Bai, Richer convolutional features for edge detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2017, pp. 5872–5881, doi:10.1109/CVPR.2017.622. http://arxiv.org/abs/1612.02103, http://ieeexplore.ieee.org/document/8100105/.

[23] Z. Yu, C. Feng, M. Liu, S. Ramalingam, Casenet: deep category-aware semantic edge detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1761–1770, doi:10.1109/CVPR.2017.191.

[24] L. Yun, C. Ming-Ming, B. Jiawang, Z. Le, J. Peng-Tao, C. Yang, Semantic Edge Detection with Diverse Deep Supervision, CoRR (2018). http://arxiv.org/abs/1804.02864.

[25] P. Sangkloy, N. Burnell, C. Ham, J. Hays, The sketchy database: learning to retrieve badly drawn bunnies, ACM Trans. Gr. 35 (4) (2016) 119:1–119:12, doi:10.1145/2897824.2925954.

[26] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: Z. Ghahramani, M. Welling, C. Cortes, N.D. Lawrence, K.Q. Weinberger (Eds.), Advances in Neural Information Processing Systems, 27, Curran Associates, Inc., 2014, pp. 2672–2680. http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf.

[27] P. Isola, J.-Y. Zhu, T. Zhou, A.A. Efros, Image-to-image translation with conditional adversarial networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2017, pp. 5967–5976, doi:10.1109/CVPR.2017.632. http://arxiv.org/abs/1611.07004 http://ieeexplore.ieee.org/document/8100115/.

[28] Z. Yi, H. Zhang, P. Tan, M. Gong, DualGAN: unsupervised dual learning for image-to-image translation, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), IEEE, 2017, pp. 2868–2876, doi:10.1109/ICCV.2017.310. http://arxiv.org/abs/1704.02510 http://ieeexplore.ieee.org/document/8237572/.

[29] M. Arjovsky, S. Chintala, L. Bottou, in: Wasserstein GAN, 2017. http://adsabs.harvard.edu/abs/2017arXiv170107875A Provided by the SAO/NASA Astrophysics Data System.

[30] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: towards real-time object detection with region proposal networks, in: Proceedings of the Twenty-Eight International Conference on Neural Information Processing Systems, in: NIPS'15, 1, MIT Press, Cambridge, MA, USA, 2015, pp. 91–99. http://dl.acm.org/citation.cfm?id=2969239.2969250.

[31] M. Mirza, S. Osindero, Conditional generative adversarial nets, CoRR abs/1411.1784 (2014). http://arxiv.org/abs/1411.1784.

[32] A. Radford, L. Metz, S. Chintala, Unsupervised representation learning with deep convolutional generative adversarial networks, CoRR abs/1511.06434 (2015). http://arxiv.org/abs/1511.06434.
[33] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, P. Abbeel, Infogan: interpretable representation learning by information maximizing generative adversarial nets, in: Proceedings of the Thirtieth International Conference on Neural Information Processing Systems, in: NIPS'16, Curran Associates Inc., USA, 2016, pp. 2180–2188. http://dl.acm.org/citation.cfm?id=3157096.3157340.
[34] J.-Y. Zhu, T. Park, P. Isola, A.A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), IEEE, 2017, pp. 2242–2251, doi:10.1109/ICCV.2017.244. http://arxiv.org/abs/1703.10593 http://ieeexplore.ieee.org/document/8237506/.
[35] G. Huang, Z. Liu, L. van der Maaten, K.Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2017, pp. 2261–2269, doi:10.1109/CVPR.2017.243. http://arxiv.org/abs/1608.06993 http://ieeexplore.ieee.org/document/8099726/.
[36] C. Li, M. Wand, Precomputed real-time texture synthesis with Markovian generative adversarial networks, CoRR abs/1604.04382 (2016). http://arxiv.org/abs/1604.04382.
[37] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, A.C. Courville, Improved training of Wasserstein GANs, in: I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, 30, Curran Associates, Inc., 2017, pp. 5767–5777. https://arxiv.org/pdf/1704.00028.pdf.
[38] M.D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, in: D. Fleet, T. Pajdla, B. Schiele, T. Tuytelaars (Eds.), Proceedings of the Computer Vision – ECCV 2014, Springer International Publishing, Cham, 2014, pp. 818–833.
[39] R. Girshick, Fast r-cnn, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), in: ICCV '15, IEEE Computer Society, Washington, DC, USA, 2015, pp. 1440–1448, doi:10.1109/ICCV.2015.169. https://doi.org/10.1109/ICCV.2015.169.
[40] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S.E. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, CoRR abs/1409.4842 (2014). http://arxiv.org/abs/1409.4842.
[41] J. Canny, A computational approach to edge detection, IEEE Trans. Pattern Anal. Mach. Intell. PAMI-8 (6) (1986) 679–698, doi:10.1109/TPAMI.1986.4767851. http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4767851.
[42] Y. Qi, J. Guo, Y.-Z. Song, T. Xiang, H. Zhang, Z.-H. Tan, Im2sketch: sketch generation by unconflicted perceptual grouping, Neurocomputing 165 (2015) 338–349, doi:10.1016/j.neucom.2015.03.023. http://www.sciencedirect.com/science/article/pii/S0925231215003082.
[43] X. Zhang, X. Li, S. Ouyang, Y. Liu, Photo-to-Sketch transformation in a complex background, IEEE Access 5 (2017) 8727–8735, doi:10.1109/ACCESS.2017.2707394. http://ieeexplore.ieee.org/document/7932876/.

Xianlin Zhang received her B.E. and Master's degrees from the Qufu Normal University and the Institute of Information Engineering of Northeastern University. She is currently a Ph.D. student at the Beijing University of Posts and Telecommunications. Her research interests include sketch-based image retrieval, image pattern recognition and deep learning.

Xueming Li received his B.E. degree from the University of Science and Technology of China in 1992 and his Ph.D. degree from the Beijing University of Posts and Telecommunications in 1997, both in electronics engineering. From 1997 to 1999, he was a postdoctoral researcher at the Institute of Information Science of Beijing Jiaotong University. He has worked for BUPT since 1999. In 2002, he was a guest lecturer at Karlsruhe University, Germany. His current research interests include digital image processing, video coding and multimedia telecommunication. To date, he has undertaken many state and enterprise RD projects, and he has published three books and more than 50 papers in the field of multimedia information processing and transmission. Professor Li is a senior member of the Chinese Institute of Electrics and a senior member of the China Society of Image and Graphics.

Xuewei Li received her B.E. and Master's degrees from the Xi'an University of Technology, at which she is currently a Ph.D. student. Her research interests include machine learning and image analysis.

Mengling Shen graduated from Beijing University of Posts and Telecommunications in 2017 with a bachelor's degree with a major in digital media technology. She is currently studying for a Master's degree in BUPT. Her main research directions are machine learning and pattern recognition.