浙江工業大学

Undergraduate design specification (thesis)

(Class of 2022)

Thesis title Sketch-based fine-grained images Research and implementation of search methods

Author Name Tong Sitao

Instructor Bai Cong

Subject (Major) Software Engineering Class 1801

College School of Computer Science and Technology

Submission Date May 30, 2022

Abstract

With the advent of the Internet era, online shopping has become the choice of many people. Merchants display a large number of product images to facilitate customers' understanding of product details; customers need to describe product attributes and other related information through keywords when searching for products. However, for some items, especially those related to fashion, such as clothes and shoe styles. It is difficult to use keywords to describe clearly the characteristics of related items. If people can retrieve the corresponding products by sketching the products, it can enrich the way users find the products. Sketching is another important way to describe an object. A simple sketch with a few strokes can contain a lot of information that needs to be described by keywords. Users can use sketches to further describe their ideas, search the target gallery, and then get the items they really need. Such a technique is called hand-drawn sketch-based image retrieval.

In this paper, implements a prototype sketch-based image retrieval system that implements a retrieval function based on hand-drawn sketches. Based on the sketches input by the user at , the corresponding alternative images in the database are queried in real time, and the alternative images are sorted by and returned to the user at . In order to further improve the retrieval accuracy, this paper proposes and implements a deep self-attentive transform triplet network sketch-based image retrieval (TF-SBIR) based on the fine-grained image retrieval algorithm of . The method uses the sketch vectorized dataset, rasterizes the sketch stroke sequences and inputs them into the improved triadic neural network for training, and uses ViT (vision transformer) to replace the convolutional neural network that processes the branch content, and the results outperform the retrieval results of the traditional method and are higher than its Top1 and Top10 accuracy.

Experiments conducted on the QMUL-ShoeV2 and QMUL-ChairV2 datasets show that the TF-SBIR retrieval method outperforms traditional methods and that the system is effective in helping users to obtain physical images corresponding to sketches Results .

Keywords: sketch-based image retrieval, fashion artificial intelligence, ternary neural network, Transformer

Abstract

With the advent of the Internet era, online shopping has become the choice of many people. Merchants display a large number of product images to facilitate Merchants display a large number of product images to facilitate customers' understanding of product details; If people can retrieve the corresponding products by sketching the products, it can enrich the way users Sketching is another important way to describe an object. A simple sketch with a few strokes can contain a lot of information that needs to be described by keywords. Users can use sketches to further describe their ideas, search the target gallery, and then get the items they really need. Such a technique is called hand-drawn sketch-based image retrieval.

In this paper, a prototype sketch-based image retrieval system is implemented to realize the retrieval function based on hand-drawn sketches. Based on the information of strokes input by the user, it is transformed into a sketch of a specific size, the corresponding alternative image in the database is the information of strokes input by the user, it is transformed into a sketch of a specific size, the corresponding alternative image in the database is Based on the information of strokes input by the user, it is transformed into a sketch of a specific size, the corresponding alternative image in the database is queried in real time, and the content of the alternative image is analyzed to obtain information such as color and model of the retrieved object. improve the retrieval accuracy, this paper proposes and implements TripleFormer-Sketch Based Image Retrival (TF-SBIR), a deep self-attentive transformed triplet network based on sketch, based on fine-grained image retrieval. The method uses the sketch vectorized dataset, rasterizes the sketch stroke sequences into the improved triple-former neural network, and uses ViT (vision transformer) to replace the convolutional neural network that processes the branch content, and the results outperform the retrieval results of the traditional method, with higher accuracy than its Top1 and Top10.

Experiments conducted on QMUL-ShoeV2 and QMUL-ChairV2 datasets show that the TF-SBIR retrieval method outperforms the traditional method, and the system can effectively help users to obtain the results of sketches corresponding to physical images. The system can effectively help users to obtain the results of sketches corresponding to physical images.

Keywords: Sketch Based Image Retrival , FashionAI, Triplet Loss, Transformer

Catalog

Figure Catalog

Table of Contents

Chapter 1 Introduction

1.1 Research Background and Significance

With the development of e-commerce business on the Internet, many images of goods are uploaded on the web, where people can freely browse and use them. These data are large in quantity and high in quality, and most of them are close-ups of the products, which are easy to do further processing. With the development of computer vision supported by artificial intelligence , more fashion design styles are included in the research. Technically speaking, AI fashion research is quite a challenging problem because fashion items are different from ordinary items, and the style and design of fashion items are more important features. To help people find their favorite fashion items, many e-commerce sites support search by keywords. However, many visual features are difficult to translate into language, so researchers often need to perform cross-domain image retrieval, for example, to match images of actual fashion items to online shopping. However, there is always a cross-domain problem for fashion image retrieval. The cross-domain problem refers to the difference in lighting, angle or orientation between the actual picture taken and the item picture. If the results are obtained by comparing shallow features, the accuracy rate is low [2].

Figure 1-1 Retrieval of the same dress [2]

A very pair of researchers have conducted related studies for fashion image based content retrieval and recommendation, such as the automatic image based clothing retrieval developed by Wang and Zhang [3]. More research has been done for cross-scene clothing retrieval.Liu et al. proposed an unsupervised migration learning approach

[4] based on part alignment and reconstruction.Kalatisdis et al. proposed an approximation to human analysis for one type of clothing retrieval [5], this research is based on manually annotated features.Huang et al. developed a multiheaded attention-ranking network (Dual Attribute-aware Ranking Network) [6], and Li et al. provided a super-resolution method for retrieved images to improve retrieval accuracy [7]. To perform retrieval among items of the same type, Wang applied a twin neural network with shared weights [8]. And Yu updated the sketch-based image retrieval architecture by acquiring images of footwear items on Taobao and Amazon and organizing volunteers for sketching to construct the QUML-Shoe dataset so that it can be used to train fine-grained sketch-based image retrieval. Previous sketch-based image retrieval algorithms did not work well for fine-grained retrieval, but after that, sketch-based image retrieval can also be used for fine-grained retrieval of fashion images.

Sketch-based image retrieval is a long-established retrieval method, and in the new era, it provides an efficient way to retrieve images. Before the camera was invented, drawing was the only option in order to record and disseminate a specific image. In ancient times, there was the search by drawing, which shows that finding corresponding items based on sketches has a long historical background. With the advent of photography, drawing gradually faded out of sight, especially in today's society, where cameras are sophisticatedly integrated into smartphones and taking pictures has become more convenient. Photographs have replaced drawings, allowing for a more accurate record that can be used as a basis for later retrieval. However, this does not mean that the application scenario of sketching search has died out. There will always be occasions when people are not comfortable taking out their phones to take pictures, or when they need to reproduce an item they have seen later. This is where people have to rely on sketching to look up the corresponding items. Sketching does not necessarily require professional drawing skills, therefore, sketch-based image retrieval systems need to capture the details in sketches to more accurately reflect the characteristics of sketches and thus improve the accuracy of sketch-based image retrieval.

In addition, with the popularity of smart devices, we can more easily build sketches, for example, using a smartphone to draw on the screen, using this way, we can improve the speed of retrieval. Although finger painting on mobile devices may lose some precision and accuracy, it is a more amateur-friendly way to paint. In order to better capture finger painting information on the screen, we have developed a method that allows users to obtain more accurate results with sparse strokes.

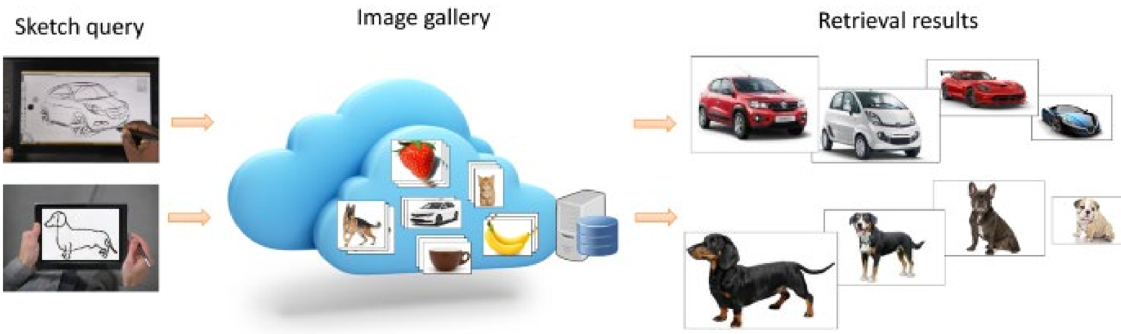

Fig. 1 1-2 way of performing sketch-based image retrieval [1]

Sketching as a more abstract representation, the cross-domain problem is more serious here. Also therefore there are many studies currently used to address the cross-domain problem in sketch-based image retrieval problem. For example, Eitz used HOG features to represent edge maps and thus unify sketch and solid map representations [9]. While Yu proposed a triadic body twin neural network, each branch network is a convolutional neural network called Sketch-a-Net that is used for fine-grained sketch-based image retrieval [10]. They developed a QMUL dataset for this task, containing ternary body annotation information, dedicated to training a ternary twin neural network [11]. Sketch generation has stimulated the creation of sketch-based image retrieval, and Ha et al. proposed SketchRNN based on recurrent neural networks that can be used to generate different sketching results [12]. Also, there are related studies that investigate the use of sketches to generate images, such as SketchyCOCO proposed by Gao. These works inspired a new way to organize strokes, i.e., strokes are stored as coordinate vectors that can preserve some other information about the pen state during drawing [13].

However, the improvement of these methods for retrieval accuracy is still limited, and the comparison by pixels and solid images is not necessarily the most human cognitive way of drawing. Meanwhile, the audience of sketch-based image retrieval should be non-professionals, and the sketches drawn are inevitably distorted to some extent. These exacerbate the level of abstraction of sketches, making the domain difference between sketches and physical images increase. It is more difficult to perform retrieval.

From the above description, we can derive the following remaining problems with the existing framework.

First, most of the current sketch-based image retrieval systems are in the experimental stage and rarely applied in practical scenarios. Taking the development of the QMUL dataset as an example, the retrieval system they implemented is only applied under the experimental environment and is only capable of drawing and querying on mobile devices, which is not a system for general users. Therefore, it is necessary to develop a browser-based sketch-based image retrieval system that can provide sketch-based image retrieval services to a wider range of users.

Second, most of the current sketch-based image retrieval algorithms still use traditional convolutional neural networks as branches, which have been able to achieve better retrieval results. However, Vision Transformer (ViT), which uses Transformer structure, has recently achieved higher results than convolutional neural networks for image classification tasks in the field of computer vision. Therefore, we hope to apply ViT to sketch-based image retrieval tasks to verify whether it can also outperform traditional convolutional neural networks for sketch-based image retrieval tasks.

1.2 Work on this paper

To meet the above requirements, the contribution of this paper can be summarized as follows:

1)Based on the need of training vectorized coordinate datasets sketch-based image retrieval models, a retrieval algorithm, the Deep Self-Attention Transformation Triad Network sketch-based image retrieval (TripleFormer-Sketch Based Image Retrival, TF-SBIR), which can construct a sequence model of images by ViT (Vision Transfor), and then obtain deeper semantic information.

2) built a sketch-based image retrieval system - retrieving the corresponding image

information in real time through the sketch drawn by the user and getting the relevant recommendation of the product, so that the user can get a better retrieval experience.

Finally the article validates the sketch TF-SBIR algorithm on QMUL-ShoeV2 and QMUL-ChairV2, and compares it with the current mainstream algorithms, proving that its has better performance than the traditional method .

1.3 Thesis Organization

This paper is divided into seven chapters, and each part is arranged as follows.

Chapter 1, introduces the research background of this topic and the main work of the topic research.

In Chapter 2, the technical background and the current state of domestic and international research related to sketching, fashion artificial intelligence and sketch-based image retrieval are introduced.

In Chapter 3, the structure of the overall system and the details of each part of the model are presented.

In Chapter 4, the experimental results are shown, and the results are analyzed and compared and discussed.

Chapter 5, shows the development of the demo system and the results of the visualization of the demo interface.

In Chapter 6, the full text is summarized and an outlook for future research is provided.

1.4 Summary of this chapter

This chapter introduces the overall research background of the project and clarifies and analyzes the significance and technical difficulties of the sketch-based image retrieval problem of . After that, the main work and the organization of this paper are introduced respectively.

Chapter 2 Overview of the Current State of Research

2.1 Fashion Artificial Intelligence

Fashion is our way of expressing ourselves to the world, and the way we dress is the basis for differentiating ourselves from others. It is therefore a unique way of expressing ourselves. The global fashion market is over $300 billion, accounting for nearly two percent of the world's GDP. Current research on AI fashion is not limited to detecting fashion items in images, but also integrates specific analysis, synthesis and recommendations for images and other content. The mainstream applications of fashion AI include fashion detection, fashion analysis and object retrieval.

With the gradual development of the e-commerce field, more and more product images are energizing the application of fashion AI. In addition to the images, the corresponding simple description of the product is also one of the materials that AI can be applied. However, the differences in angles and so on between some of the product pictures and the physical shots require a cross-domain analysis and retrieval. This constitutes a major challenge for current fashion AI. In order to solve this problem, many e-commerce platforms and Internet companies have developed various algorithms on their own to solve this problem. For example, in 2018 Ali held the Tianchi FashionAI National Challenge, which offered a prize pool of 134,000 yuan and involved more than 2,000 teams. The competition was about bite-sized annotation of the upper body image of a model for a product image, which in turn helped solve the cross-domain problem. This is a key problem in fashion AI, bit detection, and FashionNet created by Liu et al. in 2016 provides a viable approach to the process of going from physical objects to merchandise images, and they created DeepFashion, a dataset containing 800,000 images, capable of classifying and retrieving fashion images from different features [19]. In 2019 Ge et al. improved DeepFashion2 to include pose detection, making the search results more accurate [20].

Based on the above research, fashion AI can provide retrieval research ideas, while sketch-based image retrieval for apparel, shoes, clothing and hats can also be considered as a task for fashion AI.

2.2 Sketch representation and application

Sketching has been a common way of expression in people's daily life since ancient times, and with the popularity of mobile devices, it has become easier to collect and use amateurs' sketches, so related research has become a major hot topic.

The main applications of sketching are currently as follows: generates images using sketches, retrieves images or 3D models using sketches.

The main application of image generation using sketching in the early era of deep learning is generative adversarial networks (GAN), which generates images by learning the texture or content of scenes and images, and thus by sketching. The main ones in this direction are Lu et al. who designed a texture-based sketch generation technique, ContextualGAN [14]. Meanwhile, Chen et al. designed a generation method based on edge image mapping, SketchyGAN [15].Gao et al. designed an adversarial network based on vector bridging, EdgeGAN and constructed a dataset, SketchyCOCO, to evaluate the performance of EdgeGAN [16].

Methods for image retrieval using sketch-based image retrieval in the deep learning era mainly contain retrieval methods pioneered by Yu based on a ternary somatic neural network and a convolutional neural network sketch-a-Net construction, and the QMUL-ShoeV2 dataset used to train the network.Sounak et al. on the other hand proposed a zero-learning approach based on a ternary neural network Doodle to Search [17], which adds attention layer and semantic reconstruction approach after convolutional neural network to combine cross-modal textual information for retrieval.Ayan et al. on the other hand proposed a real-time sketch-based image retrieval method based on reinforcement learning, which is able to use unfinished sketch results for retrieval and improve its accuracy [18].

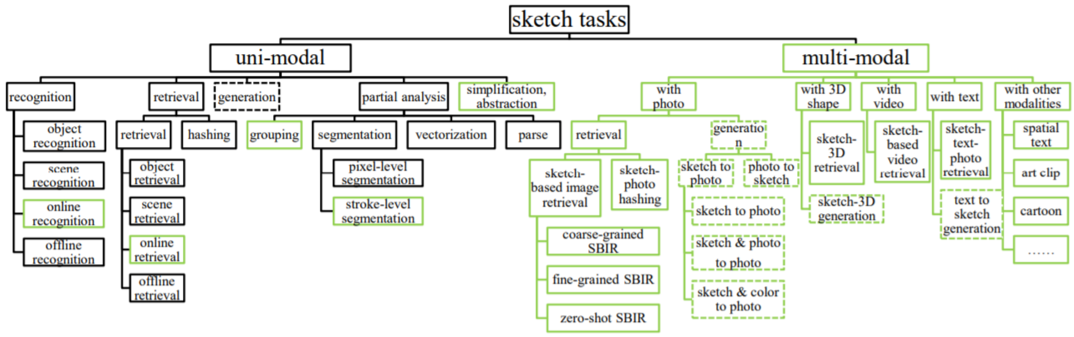

Figure 2-1 Main areas of sketch-based image retrieval [46]

Among these methods, different representations are used for sketches, and most of them use sketch images directly as input, using an end-to-end approach directly into the network. This increases the storage space for sketches, so we can use the dataset that stores the coordinate set for rasterization and then compare it with the image and sketch results.

2.3 Ternary neural network

Siamese Neural Network (SNN) has a long history of applications in comparing similarity. Many methods of comparing similarity have been applied when two vectors are compared, such as Euclidean distance, Pearson's correlation coefficient, etc. However, if this problem is applied to more complex features, these comparison methods may no longer be appropriate, in which case twin neural networks, which contain multiple unique neural networks, each of which can learn the implicit features of the input, may be the best choice. The networks can compare the inputs in parallel and thus derive the semantic similarity between the two.

We can build a three-way twin neural network, compare positive and negative examples, and continuously improve the ability to distinguish between positive and negative examples to arrive at a network that is sufficient to make judgments and thus able to perform functions such as sorting.

Among the applications of twin neural networks, the main ones are face recognition and image retrieval. twin neural network was first applied to the task of face recognition and verification by Chopra et al [21]. Paisios applied it to clothing retrieval for mobile devices [22] to compare the similarity between photographed images and clothing in a gallery. yu applied a ternary neural network to the task of fine-grained sketch-based image retrieval task and obtained good results [11].

Although ternary neural networks are widely used in the field of similarity comparison, since most of the methods are based on convolutional neural networks, while the latest ViT outperforms traditional convolutional neural network structures in image classification tasks but is not used for sketch-based image retrieval tasks, this paper proposes the TF-SBIR method, tries ViT as a branching ternary neural network, and analyzes contrasts with traditional methods and convolutional neural network based advantages and disadvantages.

2.4 Transformer

Recurrent neural network (RNN) is used to process models with sequences such as audio or natural language , which are widely used in natural language processing. To improve the performance of recurrent neural networks, introduced gated recurrent units (GRU) or long short-term memory networks (LSTM) to improve the context retention. Then, the attention mechanism was introduced to greatly improve the performance of recurrent neural networks. Later recurrent neural networks used the self-attentive mechanism, where the input senses are called values and each value is paired with a build, which can be imagined as an involuntary cue for the sensory input and thus the output. transformer, on the other hand, discarded recurrent neural networks and convolutional layers completely and was directly based on the attention mechanism, which achieved good results and was used in various fields of deep learning.

Vaswani et al. were the first to propose an encoder-decoder model containing only attention mechanisms, called Transformer [23].Xu applied Transformer to sketch recognition to generate vectorized models from original sketches [24].Leo et al. proposed a Transformer structure for encoding sketch vectors, called Sketchformer, which can be used in sketch-based image retrieval, classification, and generation [25].

The main obstacle to the application of Transformer to visual images is that if each pixel point arrangement is fed directly into the sequence model as a sequence, this input length is too large, making training difficult. So Alexey proposed a way to divide the image equally into 16×16 Alexey proposed a model that transforms images into sequences by dividing them equally into squares and feeding them into Transformer in this way, called ViT (Vision Transformer). It has achieved results over Alexnet for target detection and image classification tasks in the field of computer vision [26].

Therefore, we decided to apply ViT to our sketch-based image retrieval task to provide a solution that is superior to traditional methods.

2.5 Summary of this chapter

This chapter introduces the current state of domestic and international research on each related technology of sketch-based image retrieval at from four different aspects. Firstly, some research background and basic contents of sketch representation and application, fashion artificial intelligence are introduced respectively. After that, the technical backgrounds of triadic neural networks and Transformer are further introduced to address the core technical requirements of sketch-based image retrieval systems at .

Chapter 3 Sketch-based Image Retrieval System Architecture

3.1 Overall system analysis

The purpose of building the TF-SBIR system is to develop a sketch-based image retrieval system for users by collecting sketch images for queries. Figure 3-1 shows the training process and testing process of the system. In the training process, the system reads the contents of the dataset by batch, rasterizes the stroke coordinate data into pixel points, and then inputs them into the improved triadic neural network; the testing process of the system is to obtain the sketch input, compare the distance between the sketch result and other inputs, and then obtain the corresponding result. Therefore, the main modules of the system in the training process are sketch rasterization module and retrieval module, which will be described in detail in . The system design will be briefly described at the end.

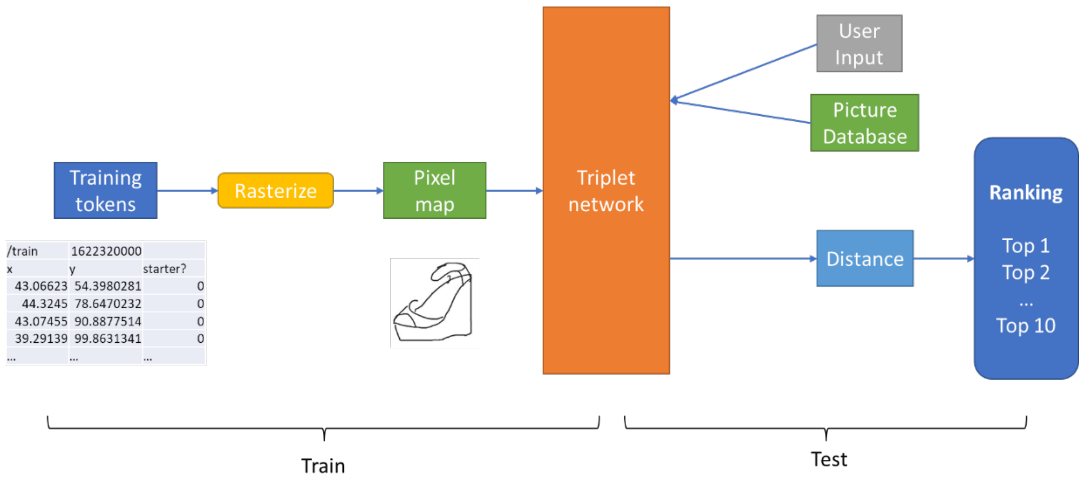

Figure 3-1 Schematic diagram of TF-SBIR system

3.2 Sketch Rasterization Module

In the training set we use here, the strokes of are stored as a three-dimensional array, which we can represent as $P_t = (x_t, y_t, q_t)$, where (x_t, y_t) denotes the coordinate points on the canvas, andq_t denotes whether it is the starting stroke or not. Ifq_t is 0, it means that this point is a continuation of the previous stroke, while ifq_t is 1, it means the beginning of a new stroke. We need to rasterize the coordinate set in this training set, and for this purpose, we use the bresenham algorithm to fill the corresponding pixel points. After collecting the coordinates of all the corresponding pixel points, we represent the rasterized image as an RGB image$\mathbb{R}^{H\times W\times 3}$ We consider all the pixel points as black, so we fill the resulting pixel points into an RGB image space of size$H = 256, W = 256$ Therefore, we fill the obtained pixel points into the RGB image space of size

The model framework structure is shown in Figure 3-2.

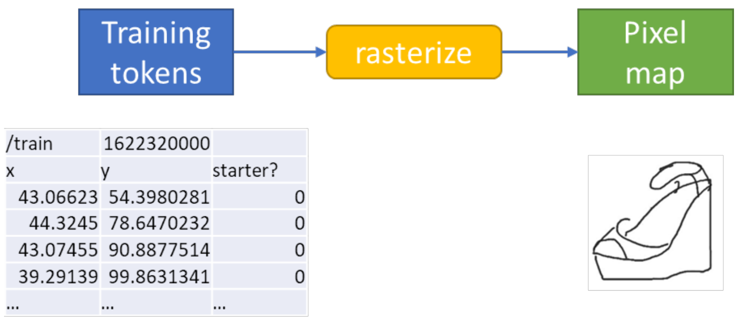

Figure 3-2 Schematic diagram of the sketch rasterization module

3.3 Sketch-based image retrieval algorithm

3.3.1 Model Overview

Our proposed TF-SBIR system is a retrieval algorithm that collects sketches and then retrieves the image library. Therefore, a relatively important approach is the need to improve the accuracy of the retrieval. After obtaining the retrieval results, some analysis of the retrieval results is needed to obtain more information about the images. Therefore, we build a Transformer-based network to extract the depth features of the sketches, which helps us to compare different images and obtain the corresponding similarity.

The idea of the TF-SBIR sketch retrieval algorithm is as follows.

Suppose there is a training set$X = \{x_{\mathrm{i}} = [s_{\mathrm{i}}, p_i], i = 1,2,\cdots,\mathrm{n}\}$, which includes a sketch set $S = [s_1, s_2 \cdots s_n] \in \mathbf{R}^{\mathrm{H\times W\times C}}$ and a set of images $P = [p_1, p_2 \cdots p_n] \in \mathbf{R}^{\mathrm{H\times W\times C}}$. Let S denote an image. For an S pair there is a corresponding image token, so that$\{s_{\mathrm{i}}, p_{\mathrm{i}}\}_{i=1}^{n}$ denotes the semantic relationship between sketches and images.

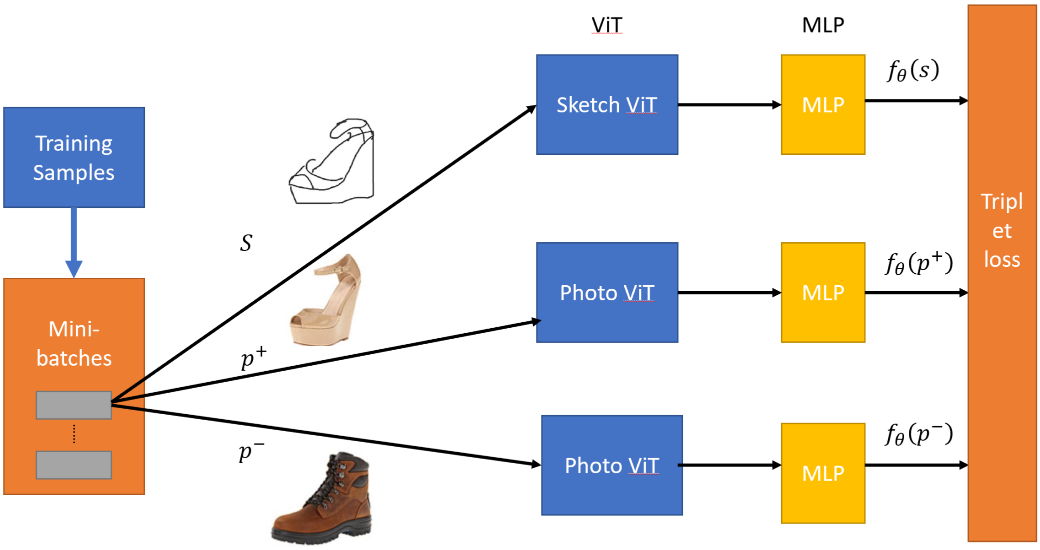

Figure 3-3 Schematic diagram of TF-SBIR model

For each given query sketch s and M alternative image libraries$\{p_j\}_{j=1}^{M} \in P$ We need to calculate the similarity between s and p and use it as a basis for ranking, we want the sketch of the query statement to rank at the top, to solve the detail gap between cross-domain and fetch images, we use a ViT based network to calculate the domain difference, we use$f_\theta(\cdot)$ to determine the$s$ and$p$ The Euclidean distance between

$$D(s,p) = ||f_\theta(s) - f_\theta(p)||_2^2$$

Then, to learn the expressions $f_\theta(\cdot)$ we used a collection of labeled triples$\{s_i, p_i^+, p_i^-\}$ as supervision. A ternary neural network is thus appropriate. In particular, each ternary neural network contains one query sketch s and two imagesp^+ andp^- called positive and negative examples, and our goal is to learn a mapping that maps the pictures and sketches to an implicit space$\mathbb{R}^d$ where the distance between the query sketch s and the positive examplep^+ is always smaller than the distance between the query sketch s and the counterexamplep^- and the counterexample:

$$D\big(f_\theta(s), f_\theta(p^+)\big) < D(f_\theta(s), f_\theta(p^-))$$

We restrict the implicit layer to the d-dimensional space, the$\big||f_\theta(\cdot)|\big|_2 = 1$

Finally, to reach this goal, we construct a deep ternary neural network and a ranking loss that uses a maximum edge error framework for a given ternary neural network$t = (s, p^+, p^-)$ its loss we define as.

$$L_\theta(t) = max\Big(0, \Delta + D\big(f_\theta(s), f_\theta(p^+)\big) - D\big(f_\theta(s), f_\theta(p^-)\big)\Big)$$

whereΔ is the edge distance of positive and negative examples, if the distance difference between the sketch and the positive and negative images is within this interval, it means that the positive and negative examples are not distinguished, and this triadic neural network will increase the loss, we optimize as follows.

$$\min_\theta \sum_{t\in T} L_\theta(t) + \lambda R(\theta)$$

where$T$ is the training set ,θ is the parameter of the depth model that defines the excitation from the input space to the implied space $R(\cdot)$ is a l_2 canonical interpreter$\big||\theta|\big|_2^2$ that minimizes the loss will reduce the distance between the positive and negative examples so that the image retrieval performance matches the sorted content. Enough ternary annotated data can eventually be trained to depth model to capture the subtle features of sketches and images for image retrieval.

As for how to map the content in the input space to the implied space, ViT is used to process sketches and images similarly. The next sections will explain exactly how ViT maps sketches and images.

3.3.2 ViT

Transformer was first applied to improve the accuracy of machine translation, and then, due to its parallel computing feature, it was trained on a large scale due to its inductive bias to be more general, which led to a richer and more general knowledge.

transformer's application to CV comes from Yang's published article proposing Vision transformer (ViT), which segments an image into several equal-sized blocks and feeds them into the attention layer for image recognition. This training method can train medium-sized images $x \in \mathbb{R}^{224\times224\times C}$ If we use a dataset that also happens to be medium-sized images $x \in \mathbb{R}^{256\times256\times C}$ A simple cropping can be done to train with ViT, so we can apply it to the task of sketch-based image retrieval at .

Figure 3-4 illustrates the general structure of the model. A classical Transformer typically receives a one-dimensional sequence of implicit bytes, and to represent a two-dimensional image, we transform the image $x \in \mathbb{R}^{H\times W\times C}$ into a flattened two-dimensional sequence $X_p \in \mathbb{R}^{N\times(P^2\cdot C)}$, and to avoid confusion with the resolution P, we use X to represent the query target image, where (H, W) is the resolution of the original image, C is the number of channels, and (P, P) refers to the resolution of each image batch, where $N = HW/P^2$ worth is the number of batches, which is also used as the length of the Transformer input sequence. This Transformer uses a fixed length D throughout the layer, so we flatten the input and input its D dimensions into a trainable linear space, which we call the linear space batch hidden layer.

In the standard Transrformer structure, a variable layer is included that has a multi-headed attention and a fully connected layer (MLP). The regularization (NL) of the layer is applied before each block and the residuals are added to the end of each module past the end.

The corresponding position hidden layer is added to the batch hidden layer in order to obtain the position information. We use a 1D position hidden layer that can be learned. There is no significant difference between choosing a 2D position hidden layer and a 1D position encoder from the ViT paper, and we use the 1D vector as the input to the encoder.

The overall structure of ViT can be mathematically represented as follows (Figure 3-4).

$$\begin{aligned}
\mathbf{z}_0 &= \left[\mathbf{x}_{\text{class}}\,; \mathbf{x}_p^1\mathbf{E}; \mathbf{x}_p^2\mathbf{E}; \cdots; \mathbf{x}_p^N\mathbf{E}\right] + \mathbf{E}_{pos}, \mathbf{E} \in \mathbb{R}^{(P^2\cdot C)\times D}, \mathbf{E}_{\text{pos}} \in \mathbb{R}^{(N+1)\times D} \\
\mathbf{z}'_\ell &= \text{MSA}\left(\text{LN}(\mathbf{z}_{\ell-1})\right) + \mathbf{z}_{\ell-1}, & \ell = 1 \ldots L \\
\mathbf{z}_\ell &= \text{MLP}\left(\text{LN}(\mathbf{z}'_\ell)\right) + \mathbf{z}'_\ell, & \ell = 1 \ldots L \\
\mathbf{y} &= \text{LN}\left(\mathbf{z}_L^0\right)
\end{aligned}$$

where x_{class} is the fully connected classification layer connected to the final x_p^i is the corresponding PATCH at the time of image division, where E refers to the fully connected layer that does the transformation for each patch, and E_{pos} denotes the location encoding information.

The above z_0 that is, the final Transformer Encoder layer input is obtained, and there are two operations among each Encoder layer, where MSA refers to the multi-headed attention, the MLP refers to the fully connected layer, and LN denotes the regularization operation.

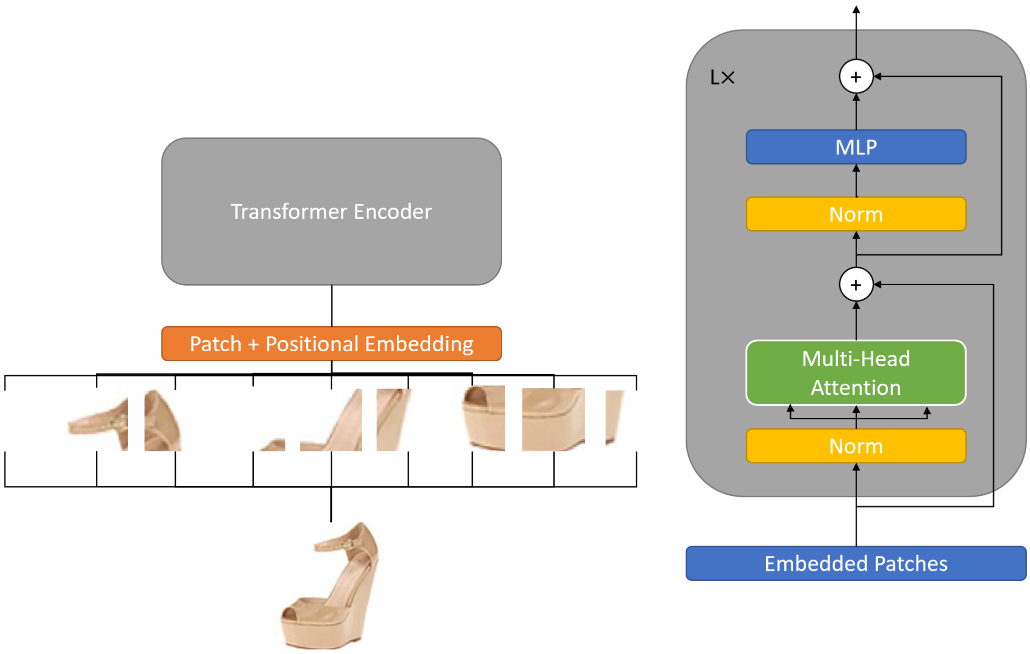

Figure 3-4 Structure of Photo ViT

3.4 Summary of this chapter

This chapter describes the specific structure of the TF-SBIR system. Firstly, the overall design requirements of the system are analyzed. This is followed by an introduction of what is required for the system and then a detailed description of TF-SBIR, the sketch-based image retrieval algorithm proposed in this paper , containing the structure of the ternary neural network and the collective description of the specific ViT network algorithm on images and sketches.

Chapter 4 Experimental results and analysis

4.1 Experimental setup

4.1.1 Experimental detail setting

The TF-SBIR algorithm is first trained using the datasets QMUL-ShoeV2 and QMUL-ChairV2, which contain 2000 training cases and 200 test cases. Among them, the query cases correspond to the set of sketch coordinate points of a particular image, while the result cases are 2000 different images. To perform small-scale data augmentation, we have a 50% probability of flipping the images or sketches when loading them.

The model is first trained on a small scale on a host computer with one Nvidia RTX 3060 graphics card. The optimizer used for the experiments was the Adam optimizer with Batch Size set to 32 and Learning Rate set to 0.0001, and 200 Epochs were trained. If the top1 accuracy is improved, we save the parameters at this point to avoid loss. The final optimal parameters are then saved separately after the training is completed.

4.1.2 Metrics

The metrics of the experiment used acc.@1, acc.@10 to react to the accuracy of the model. We defined the correctness rate (acc) as.

$$acc.@K = \frac{\sum_{q=1}^{Q} rel(q)}{Q}$$

where Q is the number of queries and K is the specified ranking. rel(q) = 1 if the correct search result exists in the first K results, and vice versa rel(q) = 0.

4.2 Sketch-based image retrieval algorithm metrics test

The experimental design is chosen to test the ternary volume dataset QMUL-ShoeV2 to evaluate the proposed sketch-based image retrieval algorithm TF-SBIR at . QMUL-ShoeV2 contains the sketch coordinate set as well as the image dataset. The experiments are performed to divide the training-validation-test set according to the setup of [1]. Each branch of the ternary neural network shares weights.

We compared the experimentally proposed algorithm with the traditional method, and the results are shown in Fig. Our method Although higher than all traditional methods on QMUL-ChairV2, it is lower on QMUL-ShoeV2 than the methods based on the bag-of-words model and the gradient direction histogram of , and lower on both datasets than the methods employing a triadic neural network structure and convolutional neural networks as branches of.

Table 5-1 Comparison of different image retrieval methods on the QMUL-ShoeV2 dataset (%)

Name	Acc.@1	Acc.@10
BoW-HOG+rankSVM [28]	17.39%	67.83%
3DS Deep + rankSVM [29]	5.22%	21.74%
Triplet+Sketch-a-Net [11]	31.3%	81.7%
TF-SBIR (ours)	**22.37%**	**66.51%**

Table 5-2 Comparison of different image retrieval methods on the QMUL-ChairV2 dataset (%)

Name	Acc.@1	Acc.@10
BoW-HOG+rankSVM [28]	28.87%	67.01%
3DS Deep + rankSVM [29]	6.19%	26.80%
Triplet+Sketch-a-Net [11]	69.07%	97.94%
TF-SBIR (ours)	**44.89%**	**88.24%**

As we can see from the table, our method has about 5% improvement in top1 over the traditional HOG and BoW based methods on the ChairV2 model, and about 17% improvement in accuracy, while performing slightly lower than this method on the ShoeV2 dataset, but higher than the 3DS Deep method.

We also need to test the performance of the experimental algorithm, which is used to test the time consumed during the whole process of sketch-based image retrieval at in the application scenario, compared with the previous bag-of-words model and the original triadic neural network, with the following results.

Table 5-3 Comparison of retrieval times of different systems on the QMUL-ShoeV2 dataset (%)

Name	Dataset size	Retrieval time
BoW-HOG+rankSVM [28]	1.5M	6s
Triplet+Sketch-a-Net [11]	~100	1-5s
TF-SBIR (ours)	**~200**	**3-4 s**

It is difficult to compare the traditional algorithms based on HoG and BoW because the datasets are not the same as our retrieval system, however and the dataset is much smaller than our system with Triplet+Sketch-a-Net network, our retrieval time can prevail. Therefore it can be concluded that our sketch-based image retrieval algorithm has some computational advantages.

4.3 Analysis of experimental results

The TF-SBIR system proposed in this project uses the original retrieval algorithm TF-SBIR at the core sketch-based image retrieval model, developed based on the latest ViT model, which is capable of mining the specific information inside the image and using images from different regions as sequence inputs to achieve parsing and matching for image content. The experimentally proposed results are higher than the traditional learning method, but the performance is still lower when comparing the ternary neural network using the improved convolutional neural network as the basis of the mapping method, so we compare the training images of both as shown in Fig.

There are several possible reasons why ViT lags behind the convolutional neural network results for the sketch-based image retrieval problem at .

First, ViT lacks Inductive Bias of Convolutional Neural Network (CNN), including Locality and Translation Equivariance). Locality refers to the fact that convolutional neural networks mainly use sliding windows to acquire features because the features of image data in adjacent regions are naturally correlated, so convolutional neural networks can retain such features. Panning invariance locality means that whether the image is panned or convolved first does not affect the result, which has the advantage that no matter where the object has moved, its output is consistent if it encounters the same convolution kernel. These inductive biases allow the CNN to have a lot of bias information, i.e., it can still get a good model using less data for learning [26].ViT does not have the above inductive biases for the location information of image blocks, so much of the scene information between image blocks needs to be relearned, and therefore, training ViT on small and medium-sized datasets cannot be compared to CNN. Our QMUL-ShoeV2 happens to be a smaller dataset that cannot provide the large amount of training data needed for ViT, and despite our data augmentation using some randomization, it still cannot meet the training requirements of ViT.

Secondly, ViT does not perform well on sparse images such as sketches in other experiments. Note that the datasets on the paperswithcode website where ViT performs well for image classification problems, such as CIRAR-10 and ImageNet, are based on actual images with context, and ViT can rank first on these datasets. For fine-grained image classification problems, ViT can only rank 23rd on datasets such as Oxford 102 Flowers. it is assumed that the Transformer model focuses more on global features and is therefore more suitable for inter-class classification problems.

The above reasons are the reasons why ViT lags behind CNN on the sketch-based image retrieval task at . To improve ViT's performance, we can target improvements for the above features. In the future, larger datasets, such as TU-Berlin and Quick Draw, can be used to perform large-scale pre-training first. Or use more dataset augmentation methods, such as stroke morphing and random removal, which in turn increase the size of the dataset.

4.4 Summary of this chapter

The main content of this chapter is conducted to compare and discuss the experimental design, experimental results and experimental results. Firstly, the settings of hyperparameters in the experiments are introduced, including experimental settings such as learning rate and batch size. After that, the detailed comparison of retrieval results between TF-SBIR and classical methods is presented. Finally shows the results of the experiment and discusses the deeper reasons and justifications for the experimental results.

Chapter 5 Demo System Development

5.1 Requirements Analysis

In order to demonstrate the full retrieval process of the TF-SBIR system in an intuitive and clear way, a demonstration system with interactive functionality was developed to show the results based on the previously trained model .

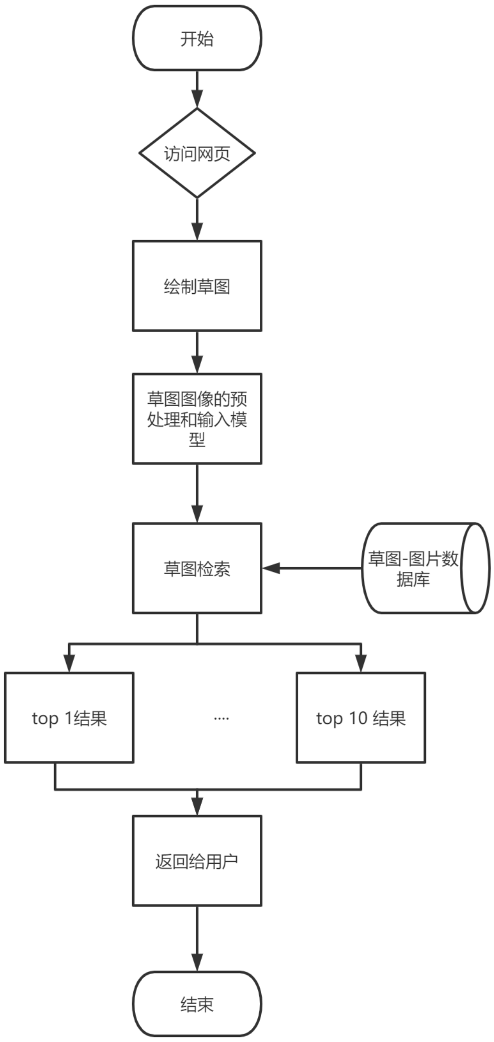

Figure 6-1 Demonstration system flow chart

In response to the above requirements, the development goal is to build a demonstration system that can provide users with sketch-based image retrieval at .

The input of the system is a hand-drawn image on the web side, and the output is a list of the top ten images of results. The operation process of the whole demo system and the related principle pictures are shown in the flow chart in Figure 6-1.

The main functions of the demo system include: according to the information of the input hand-drawn images, the size normalization process is performed, and the images are input into the ternary neural network. Then, according to the information in the image database to obtain the corresponding image similarity for ranking, to obtain the top ten images, and display the first ranked image large image.

5.2 Development Implementation

5.2.1 Development Environment

Hardware.

Intel Core i7-10350H CPU*1pc

Nvidia RTX 3060 GPU * 1pc

DDR4 ECC 8G memory 2 sticks

M2 SSD 256G one piece

M2 SSD 1T one piece

Intel AX200 Network Card

400 W power supply

Software environment.

OS: Windows 10 (WLS Ubuntu 20.04)

Programming language: python 3.8

Deep learning framework: pytorch 1.12.0

5.2.2Development Technology and Process

The development of the demo system uses the technology of front and back-end separation. The front-end graphical interface and canvas are developed with HTML5 and React.js, while the back-end uses Python and Flask framework as the back-end.

Since we need to get the hand-drawn image input by user first, we need to update the content in HTML Canvas in real time based on javascript, we designed the attached function keys needed for sketchpad, and adopted the components needed for drawing provided by Literally Canvas, so that we can do real-time drawing. Since we need to transfer Canvas images, we first store the sketch in Base64 format as a Json member , and then transfer it to the Flask server side. The server gets the information in Json format and parses the content, and finally comes up with sketch content. In the server, we call the ternary neural network to make judgment, and then sort, and return the top 10 sort results to the server, and finally return the file path corresponding to the results, and the server side parses the file path and displays the corresponding images to the web side. Therefore, the back-end mainly includes: the ternary neural network and the front-end and back-end communication modules. The Flask-based server is used to accept the input image data and is also responsible for the function of transferring the results to the front-end web page.

5.3 Results Presentation

Flask internally contains page templates, which are sent to the front-end after it has been rendered, and the page after entering the front-end is shown in Figure 6-2. The functionality of the sketchpad is briefly described as follows.

Figure 6-2 Brush function of the sketchpad

Users can select the brush color, brush size and background fill color for sketching from the above panels. Clicking on the brush color brings up the color panel to select a specific color, while dragging the slider bar can further change the brush size. When it's time to refresh the whole panel, we click clear to clear all strokes, and at the same time, users can undo and redo each step of the brush, and these actions will be reflected on the panel in real time at .

Click on the brush on the page and drag the mouse to draw the picture. After the drawing is finished, the user can click on the upload button, at which point the image information is transferred to the backend at .

Figure 6-3 Sketching board interface

After the image content is transmitted to the back-end, we judge the images in the gallery based on the trained ternary neural network, and calculate the image similarity, and then get a sorting result and send the top 10 to the front-end, as shown in Figure 6-3, the user can get the top10 and top1 results of the retrieval results of the big picture. Among them, all the images are arranged in order of retrieval, from top to bottom, where the first one is the top1 result, the second one is the top2, and so on until the last result.

Figure 6-4 Demonstration of search results

5.4 Summary of this chapter

This chapter presents the development process and results of the demo system. Firstly, the requirements of the front and back-end development of the system demonstration are analyzed, and then a demonstration of the sketch-based image retrieval based on the developed GUI program is performed to simulate the effect of the TF-SBIR system operation.

Chapter 6 Summary

6.1 Completed work

In this paper, a system framework of TF-SBIR is proposed. It is a sketch-based image retrieval system that obtains the corresponding image retrieval results based on the sketch drawing data provided by the user. Since the system needs to obtain sketches in a limited data set and to improve the efficiency of sketch-based image retrieval, this paper proposes TF-SBIR, a sketch-based image retrieval method based on a ternary neural network . The method improves the mapping function of each branch of the ternary neural network, and improves the data set storage method . The experimental results show that the system method has shorter retrieval time and higher accuracy than other deep learning retrieval methods. For the proposed sketch-based image retrieval algorithm, the complete experimental results are presented, analyzed and discussed in this paper. Finally, an interactive graphical presentation system is developed for the TF-SBIR system requirements to facilitate user interaction and experience.

6.2 Problems and Next Steps

The current framework still uses file names as query keys for retrieval and does not use other techniques such as hashing to improve retrieval efficiency and storage utilization. Therefore, in the future, we can improve the storage of image sets in the dataset to further improve the retrieval efficiency, and also to enhance the dataset migration capability.

In addition, since the front-end uses the web-side approach, the display on the mobile side is not good, so you can consider to optimize the display of on the mobile side and thus improve the user experience.

In addition, the existing model still has room for improvement, but the improvement is limited by the data set. In addition to expanding the data, we can collect users' feedback on the retrieval results during the actual application to improve the sorting of retrieval results, and then gradually improve the retrieval accuracy.

References

[1] Li Y, Li W. A survey of sketch-based image retrieval[J]. Machine Vision and Applications, 2018, 29(7): 1083-1100.

[2] Cheng W H, Song S, Chen C Y, et al. Fashion meets computer vision: A survey[J]. ACM Computing Surveys (CSUR), 2021, 54(4): 1-41.

[3] Wang X, Zhang T. Clothes search in consumer photos via color matching and attribute learning[C]//Proceedings of the 19th ACM international conference on Multimedia. 2011: 1353-1356.

[4] Liu S, Song Z, Liu G, et al. Street-to-shop: Cross-scenario clothing retrieval via parts alignment and auxiliary set[C]//2012 IEEE conference on computer vision and pattern recognition. IEEE, 2012: 3330-3337.

[5] Kalantidis Y, Kennedy L, Li L J. Getting the look: clothing recognition and segmentation for automatic product suggestions in everyday photos[C]// Proceedings of the 3rd ACM conference on International conference on multimedia retrieval. 2013: 105-112.

[6] Huang J, Feris R S, Chen Q, et al. Cross-domain image retrieval with a dual attribute-aware ranking network[C]//Proceedings of the IEEE international conference on computer vision. 2015: 1062-1070.

[7] Li Z, Li Y, Tian W, et al. Cross-scenario clothing retrieval and fine-grained style recognition[C]//2016 23rd International Conference on Pattern Recognition (ICPR). IEEE, 2016: 2912-2917.

[8] Wang X, Sun Z, Zhang W, et al. Matching user photos to online products with robust deep features[C]//Proceedings of the 2016 ACM on international conference on multimedia retrieval. 2016: 7-14.

[9] Eitz M, Hildebrand K, Boubekeur T, et al. A descriptor for large scale image retrieval based on sketched feature lines[J]. SBIM, 2009, 9: 29-36.

[10] Yu Q, Yang Y, Song Y Z, et al. Sketch-a-net that beats humans[J]. arXiv preprint arXiv:1501.07873, 2015.

[11] Yu Q, Liu F, Song Y Z, et al. Sketch me that shoe[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 799-807.

[12] Ha D, Eck D. A neural representation of sketch drawings[J]. arXiv preprint arXiv:1704.03477, 2017.

[13] Gao C, Liu Q, Xu Q, et al. Sketchycoco: Image generation from freehand scene sketches[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 5174-5183.

[14] Lu Y, Wu S, Tai Y W, et al. Image generation from sketch constraint using contextual gan[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 205-220.

[15] Chen W, Hays J. Sketchygan: Towards diverse and realistic sketch to image synthesis[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 9416-9425.

[16] Gao C, Liu Q, Xu Q, et al. Sketchycoco: Image generation from freehand scene sketches[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 5174-5183.

[17] Dey S, Riba P, Dutta A, et al. Doodle to search: practical zero-shot sketch-based image retrieval[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 2179-2188.

[18] Bhunia A K, Yang Y, Hospedales T M, et al. Sketch less for more: on-the-fly fine-grained sketch-based image retrieval[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 9779-9788.

[19] Liu Z, Luo P, Qiu S, et al. Deepfashion: powering robust clothes recognition and retrieval with rich annotations[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 1096-1104.

[20] Ge Y, Zhang R, Wang X, et al. Deepfashion2: A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 5337-5345.

[21] Chopra S, Hadsell R, LeCun Y. Learning a similarity metric discriminatively, with

application to face verification[C]//2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). IEEE, 2005, 1: 539-546.

[22] Paisios N, Subramanian L, Rubinsteyn A. Choosing which clothes to wear confidently: a tool for pattern matching[C]//Workshop on Frontiers in Accessibility for Pervasive Computing. acm. 2012.

[23] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.

[24] Xu P, Joshi C K, Bresson X. Multigraph transformer for free-hand sketch recognition[J]. IEEE Transactions on Neural Networks and Learning Systems, 2021.

[25] Ribeiro L S F, Bui T, Collomosse J, et al. Sketchformer: Transformer-based representation for sketched structure[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 14153-14162.

[26] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[J]. arXiv preprint arXiv:2010.11929, 2020.

[27] Eitz M, Hildebrand K, Boubekeur T, et al. Sketch-based image retrieval: Benchmark and bag-of-features descriptors[J]. IEEE transactions on visualization and computer graphics, 2010, 17(11): 1624-1636.

[28] Li Y, Hospedales T M, Song Y Z, et al. Free-hand sketch recognition by multi-kernel feature learning[J]. Computer Vision and Image Understanding, 2015, 137: 1-11.

[29] Wang F, Kang L, Li Y. Sketch-based 3d shape retrieval using convolutional neural networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 1875-1883.

Acknowledgements

It is hard to think that my four-year undergraduate career in college has come to an end so soon. Looking back, the road I have traveled is not a straight one. But it all came too soon, too fast. I am a person who is naturally not a tosser, but has dreams and beliefs that have made my undergraduate career full of variables. Learning is something I love, bar none, and I am able to get joy out of it, which is a gift. Although I often struggled, I never gave up. But there is nothing that can be achieved by love alone. The most important thing to thank is the people who accompanied me along the way.

First of all, I would like to express my gratitude to my supervisor, Mr. Bai Cong, for the scientific enlightenment I received during the one and a half years of study under him at ZIT. Mr. Bai has a forward-looking vision of his field and can precisely lead his students to the right path. I thought that most people do research as a job, a task, and that people in the lab are always just passing through. However, in Mr. Bai's lab, the seniors are always helpful and lively, which is a rare bright spot in the dull daily work. Mr. Bai's affability also allows the students in his lab to mingle with him. Mr. Bai's research output also comes from the free and efficient research environment he has created, where each student has his or her own place, which can be quite rare.

Secondly, I would like to thank my parents, they have paid a lot for my education, for my future development, they always spare no effort to support my every decision, from mechanical college to computer college, the hardships along the way I can only share with my family. My hometown is far away day by day, and I hope they are in good health as I spend less days with my parents.

Finally, I would like to thank all the teachers at ZIT, including our counselor Mr. Wang Haigui, who took care of most of our affairs, Director Tian Xianzhong of the School of Computer Science, the faculty secretary Ms. Li Yan Yan, and our homeroom teacher Mr. Li Qu. There were also my roommate brothers and all the seniors in the lab.

Giving my thanks to those who have helped me, who have never asked for anything in return, I can only pass on this helping heart to the next person.

Appendix

Annex 1 Literature Review of Graduation Design

Annex 2 Graduation design opening report

Annex 3 Foreign language translation of graduation design (Chinese translation and original foreign language)