Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 5577735, 14 pages
https://doi.org/10.1155/2021/5577735

Research Article

A New Algorithm for Sketch-Based Fashion Image Retrieval Based on Cross-Domain Transformation

Haopeng Lei,[1] Simin Chen,[1] Mingwen Wang,[1] Xiangjian He,[2] Wenjing Jia,[2] and Sibo Li[1]

[1]*School of Computer and Information Engineering, Jiangxi Normal University, Nanchang 330022, China*
[2]*School of Electrical and Data Engineering, University of Technology Sydney, Sydney NSW 2007, Australia*

Correspondence should be addressed to Simin Chen; simin_chen@jxnu.edu.cn

Received 6 January 2021; Revised 4 April 2021; Accepted 24 April 2021; Published 25 May 2021

Academic Editor: Amr Tolba

Due to the rise of e-commerce platforms, online shopping has become a trend. However, the current mainstream retrieval methods are still limited to using text or exemplar images as input. For huge commodity databases, it remains a long-standing unsolved problem for users to find the interested products quickly. Different from the traditional text-based and exemplar-based image retrieval techniques, sketch-based image retrieval (SBIR) provides a more intuitive and natural way for users to specify their search need. Due to the large cross-domain discrepancy between the free-hand sketch and fashion images, retrieving fashion images by sketches is a significantly challenging task. In this work, we propose a new algorithm for sketch-based fashion image retrieval based on cross-domain transformation. In our approach, the sketch and photo are first transformed into the same domain. Then, the sketch domain similarity and the photo domain similarity are calculated, respectively, and fused to improve the retrieval accuracy of fashion images. Moreover, the existing fashion image datasets mostly contain photos only and rarely contain the sketch-photo pairs. Thus, we contribute a fine-grained sketch-based fashion image retrieval dataset, which includes 36,074 sketch-photo pairs. Specifically, when retrieving on our Fashion Image dataset, the accuracy of our model ranks the correct match at the top-1 which is 96.6%, 92.1%, 91.0%, and 90.5% for clothes, pants, skirts, and shoes, respectively. Extensive experiments conducted on our dataset and two fine-grained instance-level datasets, i.e., QMUL-shoes and QMUL-chairs, show that our model has achieved a better performance than other existing methods.

1. Introduction

In recent years, the issue of fashion image retrieval has attracted increasing attention. Many research works have been reported on the tasks of clothing recognition [1, 2], clothing classification [3], and clothing retrieval [4, 5] due to their huge potential value to all walks of life. When consumers search for fashion images in online stores, mainstream retrieval methods are constrained by using text or example images as input. Due to the limited key words provided by online shopping platforms, it is difficult for consumers to retrieve the interested fashion image from the massive commodities by using text-based fashion image retrieval methods, while research on exemplar-based retrieval, where users provide an example image as the query, has recently received lots of interest in the community. However, the example images uploaded by users often suffer some problems during the actual retrieval process, such as poor light, posture changes, different shooting angles, and other factors. It is impractical to require users provide ideal example images as query input, which makes the fashion image retrieval even more challenging. A fast and effective fashion image retrieval method is currently the most urgent need for users.

Meanwhile, the way of human-computer interaction has changed dramatically due to the popularity of electronic devices. The way that humans use to retrieve fashion images is no longer restricted to using text and example images. Instead, people can use fashion sketches drawn on a touchscreen as input. For a long time, sketch is a general form of communication. Using sketch as input for retrieval has the following four advantages: (1) Fashion sketch contains more content than text does; (2) sketch is highly illustrative; (3) sketch is easy to express the styles of fashion image without

any ambiguity; (4) compared with example image, fashion sketch is easier to obtain; etc. Recently, the research related to sketch has flourished. Up to now, many problems have been studied, including sketch recognition [6, 7], sketch-based image retrieval (SBIR) [8, 9], and sketch-based 3D model retrieval [10], just to name a few. What is more, sketch-based fashion image retrieval is still relatively new. As the result, the urgent needs of users and the advantages of sketch-based retrieval provide us with a strong motivation to propose a more effective sketch-based image retrieval method, which uses sketch images as query input for fashion image retrieval.

With the above strong motivation, using sketch as input for fashion image retrieval faces, these problems to be solved in this paper. (1) Fashion sketches and fashion photos belong to two different domains. Compared with photos, sketches are composed of black lines on white background and look more abstract and lack information such as patterns, materials, and colours. This unique characteristic of the sketches increases the difficulty of fine-grained fashion image retrieval. (2) Most of the existing fashion image retrieval methods are based on example images input query. Images having similar visual content will be returned to the users by calculating the similarity between query image and database images. However, the input example images often have problems such as poor light, posture changes, different shooting angles, and complex background, which make it difficult to retrieve specific styles of fashion image for the users. (3) It is very difficult to collect fashion sketches. To the best of our knowledge, there is no large-scale dataset available for researchers to develop advanced solutions. In addition, we will need thousands of pairs of matching fashion sketches and images for our cross-domain deep learning. Therefore, it is challenging to create such this database covering different fashion image categories.

In this work, aiming to solve the problem of sketch-based fashion image retrieval, i.e., given a sketch of a fashion product, match it with the fashion photo in the dataset, and return the true-match fashion photo, we propose an efficient and reliable framework for fine-grained sketch-based fashion image retrieval to address these challenges. The framework of our method consists of three modules, including a cross-domain transformation module, a cross-domain feature extraction module, and a cross-domain similarity measurement module. We first use the cross-domain transformation module to transform sketches and photos into the same domain, and then, we adopt cross-domain feature extraction module to extract deep features of the query fashion sketch and the fashion photos in the retrieval dataset from the sketch domain and the photo domain, respectively. Next, we calculate the similarity between the transformed photo and fashion photos in photo domain and the similarity between the query sketch and transformed sketches in sketch domain. Finally, we fuse the two similarities in the different domains to achieve the final retrieval results.

The main contributions of this work are threefold:

(1) We propose a new algorithm for sketch-based fashion image retrieval based on cross-domain transformation, which transforms the fashion sketch and the fashion photo into the same domain before retrieval. Our proposed approach eliminates the requirement of rich annotation for the dataset and solves the heterogeneous problem of fashion sketches and fashion photos. In particular, the approach can effectively improve the retrieval accuracy of fashion image

(2) Most of the existing fashion image retrieval methods are based on example images input query. Images having similar visual content will be returned to the users by calculating the similarity between query image and database images. This method only calculates the similarity of the photo domain once. While we are doing cross-domain fashion image retrieval on two domains, we first transform the query fashion sketch into a fashion photo, use the transformed fashion photo to retrieve the fashion image dataset, and perform a similarity calculation of the photo domain. And then, we transformed all the fashion photos in the dataset into fashion sketches, use the query fashion sketch to retrieve the transformed fashion sketch dataset, and perform a similarity calculation of the sketch domain. Finally, we fuse the two similarities of the photo domain and the sketch domain to calculate the final similarity to obtain a more accurate retrieval result

(3) We contribute a new fine-grained sketch-based fashion image retrieval dataset, which contains 36,074 sketch-photo pairs covering 26 fashion types. As far as we know, it is the first comprehensive sketch-based fashion image retrieval dataset.

2. Related Work

2.1. Category-Level SBIR and Fine-Grained SBIR. Category-level sketch-based image retrieval (category-level SBIR) is conventional sketch-based image retrieval. It mainly focuses on retrieving images of the same category rather than the differences of intracategory. In recent years, the problem of category-level sketch for image retrieval [11–14] has been well studied. Most of the existing methods [11–13] first learn the common feature space of the sketch and the original image, perform similarity calculation and matching, and then retrieve the object that matches the target and return the object. However, with this method of learning, the common feature space between the sketch and the image may cause the model to collapse and therefore cannot achieve the expected results.

Fine-grained sketch-based image retrieval (fine-grained SBIR) is a new concept [8, 15–17]. The first attempt to solve the fine-grained SBIR was made by Li et al. [15], which mainly applied the deformable part-based model (DPM) to SBIR. Their definition of fine-grained emphasizes the viewpoint and observation of the object depicted by the sketch. As its consequence, an ideal recall image is the one that has a posture or perspective similar to the query sketch, regardless of whether the recalled image contains the same object.

However, it is very different from ours. Our definition of fine-grained is the same as that described in [8, 18], which emphasizes the details of the object depicted in the sketch. That is to say, for a retrieved image to match the query sketch, it must contain the same object instances. In recent years, with the development of artificial intelligence technology, CNN has significantly improved the performance in various computer vision tasks, such as image classification [19], image annotation [20], image retrieval [21–23], and medical image analysis [24, 25]. Khanday and Sofi [26] reviewed the state-of-the-art technology in computer vision by highlighting the contributions, challenges, and applications. In addition, the CNN-based feature extraction also demonstrates the excellent performance in sketch-based image retrieval, i.e., in 2015, Yu et al. [27] first abandoned the traditional feature extraction method of using convolutional neural networks and proposed a sketch-a-net network structure specially designed for free-hand sketch, which performed better than that proposed by Li et al. [18]. For example, when users search for a skirt, category-level SBIR can return a series of pictures of skirts for users, which is more complex than the way users input the text "skirt" instead of drawing the appearance of the skirt, whereas fine-grained SBIR can return the skirt corresponding to these details according to the sketch details entered by the user.

2.2. Fashion Image Datasets. Since the collection of sketches is not as easy as collecting photos, a significant obstacle to the research of sketch-based fashion image retrieval is the lack of benchmark datasets. As summarized in Table 1, the existing fashion image datasets all have different shapes and sizes and can be grouped according to single vs. multimodal. The single-modal datasets only consist of fashion photos, which are mainly used for fashion image recognition and retrieval from photo to photo. Moreover, most of the fashion photos contained in these single-modal datasets have complex backgrounds. Multimodal datasets support cross-domain tasks by providing sketches and photos. For example, the QMUL-shoes dataset [8] contains 419 sketch-photo pairs of shoes. The dataset contains simple images, but the only category is shoes. So, for the fashion category, the dataset is incomplete, and the size is small. Instead, our dataset has 36,074 fashion sketch-photo pairs, including clothes, pants, skirts, and shoes, covering almost all fashion categories. Compared with the QMUL-shoes dataset, it has more sketch-photo pairs and more comprehensive coverage of fashion image categories. Some example images in different datasets are shown in Figure 1. As it shows, photos in our dataset are as simple as those in QMUL-shoes.

2.3. Generative Adversarial Networks. Generative adversarial networks (GANs) [29] have made remarkable achievements in computer vision. A GAN model typically consists of two modules, i.e., the generator G and the discriminator D. In order to fool the discriminator, the generator should learn to generate false images that are indistinguishable from real images; meanwhile, the discriminator should learn to distinguish between real images and false images generated by the generator. The learning of GAN is a zero-sum game. The final result of the game is that, under ideal conditions, it is difficult for the discriminator to judge whether the image generated by the generator is real or false, that is, $D(G(z)) = 0.5$, where z is random noise.

Table 1: The comparison of our dataset with existing datasets.

Single-modal datasets	Number of images
DDAN [28]	341,021 photos
WTBI [4]	78,958 photos
DeepFashion [2]	Over 800,000 photos
Multi-modal datasets	**Number of images**
QMUL-shoe [8]	419 sketches, 419 photos
Our Fashion Image dataset	36,074 sketches, 36,074 photos

Since sketch and photo are heterogeneous, in order to overcome this challenge, GANs are used to eliminate the domain gap. The standard GAN is a one-way generation model that requires paired training data, i.e., all sketches in the sketch domain are converted to the same photo in the natural photo domain. To eliminate this requirement, Zhu et al. [30] proposed a cycle-consistency loss and CycleGAN. CycleGAN is a bidirectional generation model that can transform the sketch into the photo domain, and then back to the sketch domain, and can work in the absence of paired examples. Inspired by this approach, in this paper, we propose to transform images' domain by enforcing the cycle-consistency constraint. The backbone framework of our proposed model is based on UNIT [31] and VGG-16 [32]. We utilize the UNIT model to transform images' domain, where the UNIT model implies the cycle-consistency constraint, which can achieve perfect conversion between the images in different domains. Then, we use the VGG-16 network till the last convolutional layer to obtain the feature vectors, and then, we measure the similarity of feature vectors. Finally, the most similar photo is returned.

3. Proposed Method

3.1. Overview. In this section, we mainly describe the collection process of the Fashion Image dataset and the retrieval process of our proposed method. The framework of our method consists of three modules, including cross-domain transformation module, cross-domain feature extraction module, and cross-domain similarity measurement module. An overview of our proposed sketch-based fashion image retrieval model based on cross-domain transformation is illustrated in Figure 2.

Given a query fashion sketch s_q and the photos $p_n (n = 1, 2, \cdots, N)$ of the dataset, where N is the total number of fashion photos in the dataset, the aim is to retrieve the true-match fashion photo of the query sketch from the dataset. The retrieval procedure of our proposed method is divided into two streams: the sketch-based fashion photo retrieval stream and the sketch-based fashion sketch retrieval stream.

3.1.1. Sketch-Based Fashion Photo Retrieval Stream. First, in order to bridge the domain gap between the sketch and the

Figure 1: Examples of the images in different datasets showing their different styles.

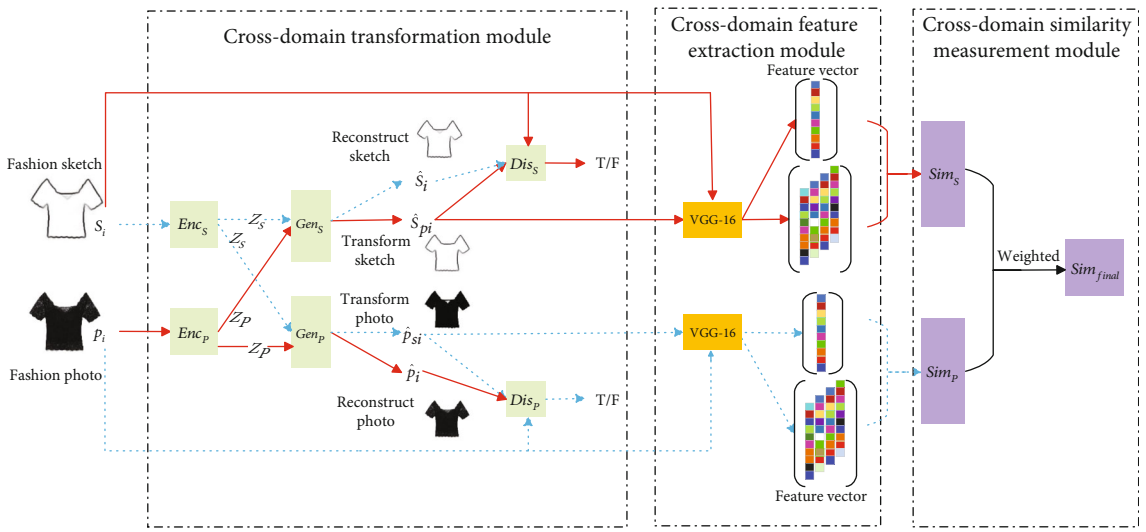

Figure 2: An overview of our proposed sketch-based fashion image retrieval model based on cross-domain transformation. Enc_S and Enc_P are encoders. Gen_S and Gen_P are generators. Dis_S and Dis_P are discriminator. The red solid arrows represent the sketch-based fashion sketch retrieval stream, and the blue dotted arrows represent the sketch-based fashion photo retrieval stream.

photo, the query sketch s_q first needs to be transformed into a fashion photo $\hat{p}_{s_q}$. Second, we extract the deep features of the transformed photo $\hat{p}_{s_q}$ and the fashion photos $p_n(n = 1, 2, \cdots, N)$ in the dataset through the cross-domain feature extraction module, respectively. Third, according to the obtained deep features, we calculate the similarity Sim_P between the transformed photo $\hat{p}_{s_q}$ and the fashion photos $p_n(n = 1, 2, \cdots, N)$. All the processes are shown by the dotted arrows in Figure 2.

3.1.2. Sketch-Based Fashion Sketch Retrieval Stream. Similar to the sketch-based fashion photo retrieval stream, we first map the fashion photos $p_n\,(n = 1, 2, \cdots, N)$ to the corresponding sketches $\hat{s}_{p_n}$. Second, we use the cross-domain

feature extraction module to extract the deep features of query sketch s_q and transformed sketches $\hat{s}_{p_n}(n = 1, 2, \cdots, N)$. Third, we calculate the similarity Sim_S between s_q and $\hat{s}_{p_n}(n = 1, 2, \cdots, N)$. All the processes are shown by the solid arrows in Figure 2.

After performing these two streams, for the query sketch s_q, we achieve the similarity Sim_P between the transformed photo $\hat{p}_{s_q}$ and the fashion photos $p_n\,(n = 1, 2, \cdots, N)$. Moreover, we also achieve the similarity Sim_S between the query sketch s_q and the transformed sketches $\hat{s}_{p_n}(n = 1, 2, \cdots, N)$. Then, we combine the sketch-based fashion photo retrieval stream and the sketch-based fashion sketch retrieval stream to improve the retrieval accuracy. Thus, we assign weights to these two similarities and add them to calculate the final similarity $\text{Sim}_{\text{final}}$. We rank the similarity $\text{Sim}_{\text{final}}$ to obtain an index table of the similarity between the query sketch s_q and the fashion photos $p_n\,(n = 1, 2, \cdots, N)$ in the dataset. Finally, according to the index table, the fashion photo which is the most similar to the query fashion sketch s_q in the dataset is returned as the retrieval result.

3.2. Fashion Image Dataset. We contribute a fine-grained Fashion Image dataset that contains a complete range of fashion types and can be used for fashion image cross-domain retrieval. We divide fashion images into four categories, i.e., clothes, pants, skirts, and shoes, and then further divide these four categories in detail. This dataset has three advantages. Firstly, as a multimodal (sketch and photo) fashion image dataset, it has a wide range of fashion categories, including clothes, pants, skirts and shoes. Secondly, it is a fine-grained dataset, where the clothes are divided into 11 subcategories, pants into 4 subcategories, skirts into 6 subcategories, and shoes into 5 subcategories. Thirdly, compared with other datasets of the same type, its size is larger, including 36,074 sketch-photo pairs. Next, we will describe the process of data collection and processing in detail.

3.2.1. Collecting Photos. The fashion photos we collect are mainly from three online shopping websites, including Taobao, Jindong, and Amazon, and a small part are from Baidu pictures and Google pictures. For fashion image, we divide them into four categories, i.e., clothes, pants, skirts, and shoes. Since the dataset we created is a fine-grained fashion image dataset, almost all relevant subcategories are included in each major category. For example, clothes consist of 11 subcategories, including suspender vests, short coats, long coats, short sleeve T-shirts, long sleeve T-shirts, short sleeve shirts, long sleeve shirts, vest, long cotton-padded jackets, short cotton-padded jackets, and leisure hoodies, covering almost all types of clothes. Finally, 12,603 representative cloth photos have been selected. For the collection of pants, skirts, and shoes, we also include different types and styles. We selected 5,610 photos of pants, including 4 types of back-belt pants, trousers, shorts, and jumpsuit; 13,321 photos of skirts, including 6 types of long skirts, mini-skirts, long sleeve dresses, short sleeve dresses, sleeveless dresses, and back-belt skirts; and 4,540 photos of shoes covering high heels, boots, flats, slippers, and sandals.

3.2.2. Collecting Sketches. The second step is to convert the collected photos to their corresponding sketches. We use the Structured Edge Detection Toolbox [33] to handle photos and obtain the edge maps, which are similar to free-hand sketches. Furthermore, in order to make the edge maps closer to the free-hand sketches, we performed an erasing operation on the edge maps, that is, to erase unnecessary line information in the edge maps and finally get the fashion sketches.

3.3. Cross-Domain Transformation Module. Since the fashion sketch and the fashion photo come from different domains, we transform the fashion sketch and the fashion photo into the same photo and sketch domain to bridge the domain gap. We propose a cross-domain transformation module, which is composed of 6 networks, namely the fashion sketch encoder Enc_S, the fashion photo encoder Enc_P, the fashion sketch generator Gen_S, the fashion photo generator Gen_P, the fashion sketch discriminator Dis_S, and the fashion photo discriminator Dis_P. The encoders include 3 convolutional layers and 4 residual basic blocks, which are used to encode the fashion sketch/photo into the latent code z_S/z_P. The generators include 4 residual basic blocks and 3 convolutional layers, which are used to decode the latent code z_S/z_P and generate the transformed fashion sketch/photo. The discriminators include 6 convolutional layers which are used to distinguish between the real fashion sketch/photo and the transformed fashion sketch/photo. The function of cross-domain transformation module includes self-reconstruction of the intradomain and the transformation of the cross-domain. We divide the cross-domain transformation module into two submodules $T_{S\to P}$ and $T_{P\to S}$. The first submodule $T_{S\to P}$ is used to transform the fashion sketch into the photo domain, and the second submodule $T_{P\to S}$ is used to transform the fashion photo into the sketch domain. The detailed cross-domain transformation training process is described as follows.

Suppose that the training sample pairs $\{s_i, p_i\}$ of the fashion sketch and the fashion photo are, respectively, given from the training dataset. We input the fashion sketch sample s_i into the first cross-domain transformation submodule $T_{S\to P}$, where the fashion sketch encoder Enc_S transforms the fashion sketch s_i into a latent code $z_S : z_S \sim \text{Enc}_S(s_i) = q_S(z_S \mid s_i)$, and the fashion sketch generator Gen_S decodes the latent code z_S to reconstruct the original input fashion sketch: $\hat{s}_i \sim \text{Gen}_S(z_S) = p_{\text{Gen}_s}(\hat{s}_i \mid z_S)$.

We use *VAE* [34–36] (variational autoencoder) to construct the encoder-decoder for the fashion sketch in our cross-domain transformation module. The objective function of the encode-decode process for the fashion sketch s_i is given by

$$\begin{aligned} L_{\text{Enc}_S} = {} & D_{\text{KL}}\left(q_S(z_S \mid s_i) \middle\| p_{\text{prior}}(z_S)\right) \\ & - \mathbb{E}_{z_S \sim q_S(z_S \mid s_i)}\left[\log\, p_{\text{Gen}_s}(\hat{s}_i \mid z_S)\right], \end{aligned} \tag{1}$$

where $q_S(z_S \mid s_i)$ represents that the fashion sketch encoder Enc_S maps the fashion sketch s_i into a latent code z_S and $p_{\text{prior}}(z_S)$ represents the prior distribution of the latent code

z_S. For simplicity, the prior distribution of latent code z_S can be assumed to follow a zero mean Gaussian distribution $N(0, I)$. $D_{\mathrm{KL}}(q_S(z_S \mid s_i) \| p_{\mathrm{prior}}(z_S))$ represents the KL divergence between the probability distribution $q_S(z_S \mid s_i)$ and $p_{\mathrm{prior}}(z_S)$. Therefore, the first term of this objective function is to ensure that the posterior distribution $q_S(z_S \mid s_i)$ of the latent code z_S is similar to the true prior distribution $p_{prior}(z_S)$. $p_{\mathrm{Gen}_s}(\hat{s}_i \mid z_S)$ represents the fashion sketch generator Gen_S that reconstruct the fashion sketch $\hat{s}_i$ given the latent code z_S. The second term of this objective function is the reconstruction loss which measures the reconstruction error between the reconstructed fashion sketch $\hat{s}_i$ and the original fashion sketch s_i.

Moreover, for the purpose of encouraging the reconstructed fashion sketch $\hat{s}_i$ to resemble the original fashion sketch s_i as closely as possible, we build the generative adversarial network GAN_S in our proposed cross-domain transformation module by combing the fashion sketch generator Gen_S and the fashion sketch discriminator Dis_S. The objective function of GAN_S is given by

$$\begin{aligned} L_{\mathrm{GAN}_S} = {} & \mathbb{E}_{s_i \sim p_{\mathrm{data}}(S)}[\log\ Dis_S(s_i)] \\ & + \mathbb{E}_{z_S \sim q_S(z_S \mid s_i)}[\log\ (1 - Dis_s(Gen_S(z_S)))], \end{aligned} \tag{2}$$

where $p_{data}(S)$ represents the probability distribution of all the fashion sketches in the training dataset. The fashion sketch generator Gen_S is used to reconstruct the fashion sketch $\hat{s}_i$ that looks similar to the original fashion sketch s_i given the latent code z_S, and the fashion sketch discriminator Dis_S is used to distinguish between the real original fashion sketch s_i and the reconstructed fashion sketch $\hat{s}_i$. Therefore, this objective function is to calculate the cross-entropy loss that encourages Gen_S to reconstruct the same original fashion sketch s_i and simultaneously provides the best discrimination ability to recognize the reconstructed sketch $\hat{s}_i$.

Then, in order to transform the fashion sketch into the photo domain, we input the latent code z_S of fashion sketch s_i into the fashion photo generator Gen_P to generate the transformed fashion photo $\hat{p}_{s_i}$, and we will input the generated fashion photo $\hat{p}_{s_i}$ and the real fashion photo p_i into the fashion photo discriminator Dis_P to determine whether an input fashion photo is the real fashion photo p_i or the transformed fashion photo $\hat{p}_{s_i}$. Fashion photo generator Gen_P and fashion photo discriminator Dis_P constitute the generative adversarial network [29] $GAN_{S \to P}$. The $GAN_{S \to P}$ objective function can be defined as

$$\begin{aligned} L_{GAN_{S \to P}} = {} & \mathbb{E}_{p_i \sim p_{data}(P)}[\log\ Dis_P(p_i)] \\ & + \mathbb{E}_{z_S \sim q_S(z_S \mid s_i)}[\log\ (1 - Dis_P(Gen_P(z_S)))], \end{aligned} \tag{3}$$

where $p_{\mathrm{data}}(P)$ represents the probability distribution of all the fashion photos in the training dataset. The fashion photo generator Gen_P tries to generate the fashion photo $\hat{p}_{s_i}$ that looks similar to the real fashion photo p_i given the latent code z_S, while fashion photo discriminator Dis_P tries to distinguish between real fashion photo p_i and the generated fashion photo $\hat{p}_{s_i}$.

Similarly, the fashion photo encoder Enc_P and the fashion photo generator Gen_P constitute a VAE network, which is used for reconstructing the fashion photos in the photo domain P. We input the fashion photo p_i into the second cross-domain transformation submodule $T_{P \to S}$. The fashion photo encoder Enc_P encodes the input fashion photo p_i into a latent code $z_P \sim Enc_P(p_i) = q_P(z_P \mid p_i)$, and the fashion photo generator Gen_P decodes the latent code z_P to reconstruct the fashion photo p_i; the self-reconstruction of the fashion photo p_i in photo domain P can be expressed as $\hat{p}_i \sim \mathrm{Gen}_P(z_P) = p_{\mathrm{Gen}_P}(\hat{p}_i \mid z_P)$. Thus, the objective function of the fashion photo p_i encode-decode process can be defined as

$$\begin{aligned} L_{\mathrm{Enc}_P} = {} & D_{\mathrm{KL}}\left(q_P(z_P \mid p_i) \middle\| p_{\mathrm{prior}}(z_P)\right) \\ & - \mathbb{E}_{z_P \sim q_P(z_P \mid p_i)}\left[\log\ p_{\mathrm{Gen}_P}(\hat{p}_i \mid z_P)\right], \end{aligned} \tag{4}$$

where the $q_P(z_P \mid p_i)$ represents the probability distribution of decoding the fashion photo p_i into the latent code z_P, the $p_{\mathrm{prior}}(z_P)$ indicates that the prior probability of the latent code z_P obeys the zero mean Gaussian distribution model $N(0, I)$, and the $p_{Gen_P}(\hat{p}_i \mid z_P)$ represents the probability distribution of the fashion photo generator Gen_P that reconstruct the latent code z_P to the fashion photo p_i. The first term is to penalize the latent code distribution that deviates from the prior distribution, and the second term is to constrain the reconstructed photo $\hat{p}_i$ to be similar to the input photo p_i.

What is more, we input the reconstructed photo $\hat{p}_i$ and the fashion photo p_i into the fashion photo discriminator Dis_P to determine whether an input fashion photo is true or false. The objective function of the generative adversarial network GAN_P composed of Gen_P and Dis_P can be defined as

$$\begin{aligned} L_{\mathrm{GAN}_P} = {} & \mathbb{E}_{p_i \sim p_{\mathrm{data}}(\mathrm{P})}[\log\ \mathrm{Dis}_P(p_i)] \\ & + \mathbb{E}_{z_P \sim q_P(z_P \mid p_i)}[\log\ (1 - \mathrm{Dis}_P(\mathrm{Gen}_P(z_P)))]. \end{aligned} \tag{5}$$

Similar to the equation (3), the fashion photo generator Gen_P is used to reconstruct the fashion photo $\hat{p}_i$ given the latent code z_P, and the fashion photo discriminator Dis_P is used to distinguish between the real original fashion photo p_i and the reconstructed fashion photo $\hat{p}_i$.

The fashion sketch generator Gen_S and the fashion sketch discriminator Dis_S constitute $\mathrm{GAN}_{P \to S}$, which is used for transforming the fashion photo p_i from the photo domain P to the sketch domain S, and the transformed fashion sketch is $\hat{s}_{p_i} = Gen_S(z_P \mid p_i)$; Dis_S is trained to distinguish between the real sketch s_i and the transformed sketch $\hat{s}_{p_i}$, which gives high scores to real sketches and

low scores to generated sketches. The $\mathrm{GAN}_{P\to S}$ objective function is given by

$$L_{GAN_{P\to S}} = \mathbb{E}_{s_i\sim p_{data}(S)}[\log\ Dis_S(s_i)] + \mathbb{E}_{z_P\sim q_P(z_P|p_i)}[\log\ (1 - Dis_S(Gen_S(z_P)))]. \tag{6}$$

At last, in order to improve the robustness and stability of the submodule $T_{S\to P}$ and $T_{P\to S}$, we need to ensure that the fashion photo $\widehat{p}_{s_i}$ transformed by the original fashion sketch s_i can be transformed back to the same sketch s_i, and the fashion sketch $\widehat{s}_{p_i}$ transformed by the original fashion photo p_i can be transformed back to the same photo p_i. Meanwhile, the original sketch features and photo features will not be lost after these twice transformation. Therefore, we utilize a cycle-consistency constraint [31] for the entire cross-domain transformation network. To achieve this goal, we input $\widehat{p}_{s_i}$ to the fashion photo encoder Enc_P for encoding and use fashion sketch generator Gen_s that decodes the latent code z_P to reconstruct the fashion sketch s_i. *VAE* can also be used to construct the encoder-decoder. The objective function of cycle-consistency constraint for fashion sketch is given by

$$L_{\mathrm{cyc}_S} = D_{\mathrm{KL}}\left(q_P\left(z_P \mid \widehat{p}_{s_i}\right)\middle\|p_{\mathrm{prior}}(z_P)\right) - \mathbb{E}_{z_P\sim q_P\left(z_P|\widehat{p}_{s_i}\right)}\left[\log\ p_{\mathrm{Gen}_s}(s_i \mid z_P)\right].$$

Similar to the above process, $\widehat{s}_{p_i}$ is input to the fashion sketch encoder Enc_S for encoding, and the fashion photo generator Gen_P is used to decode the latent code z_S to reconstruct the fashion photo p_i. The objective function of cycle-consistency constraint for fashion photo is given by

$$L_{\mathrm{cyc}_P} = D_{\mathrm{KL}}\left(q_S\left(z_S \mid \widehat{s}_{p_i}\right)\middle\|p_{\mathrm{prior}}(z_S)\right) - \mathbb{E}_{z_S\sim q_S\left(z_S|\widehat{s}_{p_i}\right)}\left[\log\ p_{\mathrm{Gen}_P}(p_i \mid z_S)\right]. \tag{8}$$

In summary, combined with equations (1), (2), (3), and (7), the total objective function of the fashion sketch cross-domain transformation submodule $T_{S\to P}$ is given by

$$L_{T_{S\to P}} = L_{\mathrm{Enc}_S} + L_{\mathrm{GAN}_S} + L_{\mathrm{GAN}_{S\to P}} + L_{\mathrm{cyc}_S}, \tag{9}$$

and combined with equation (4), (5), (6), and (8), the total objective function of the fashion photo cross-domain transformation submodule $T_{P\to S}$ is given by

$$L_{T_{P\to S}} = L_{\mathrm{Enc}_P} + L_{\mathrm{GAN}_P} + L_{\mathrm{GAN}_{P\to S}} + L_{\mathrm{cyc}_P}. \tag{10}$$

During the training process, we use the Adam optimizer to alternately optimize the objective functions $L_{T_{S\to P}}$ and $L_{T_{P\to S}}$. After the objective function optimization, the entire training process of the cross-domain transformation module can be completed. During the testing process, for any input query sketch s_q and fashion photos $p_n\ (n = 1, 2, \cdots, N)$ in the retrieval dataset, we can transform them into the same domain by using our proposed cross-domain transformation module.

3.4. Cross-Domain Feature Extraction Module. After the transformation of cross-domain images is completed, we deploy a symmetric CNN as the feature extraction module, which uses the VGG-16 [32] pretrained on ImageNet as the backbone network of the feature extraction module. If we use the pooling operation as a split point to group the entire VGG-16 network, we will get five sets of convolutions. The first two groups of convolutions have the same form, which is conv-relu-conv-relu-pool; the last three groups of convolutions have the same form, which is conv-relu-conv-relu-conv-relu-pool. In addition to the convolution group, VGG-16 has three fully connected layers at the end. However, in this paper, we use the VGG-16 network until the last convolutional layer obtains the feature matrixes. Finally, the size of feature vector obtained from VGGNet is 1×512.

For the sketch-based fashion photo retrieval stream, we use the VGG-16 network for the photo domain to extract features for photos $p_n\ (n = 1, 2, \cdots, N)$ in the Fashion Image dataset and store in the database as vectors. For the query sketch s_q, firstly, it is transformed into a fashion photo $\widehat{p}_{s_q}$, and then we use the same VGG-16 network to extract its deep feature vector of the same size. At last, we get the vector extracted from the transformed photo $\widehat{p}_{s_q}$ and the vectors extracted from photos $p_n\ (n = 1, 2, \cdots, N)$, which are used as the input of the cross-domain similarity measurement module.

For the sketch-based fashion sketch retrieval stream, photos $p_n\ (n = 1, 2, \cdots, N)$ in the Fashion Image dataset have been transformed into the corresponding sketches $\widehat{s}_{p_n}(n = 1, 2, \cdots, N)$ in the sketch domain *S*. And now, a transformed sketch dataset is obtained. We extract features for each transformed sketch $\widehat{s}_{p_n}$ by using the VGG-16 network for the sketch domain; then, the obtained feature vectors are stored in the database. We also get a deep feature vector by using the same VGG-16 network to extract features for the query sketch s_q. Finally, the feature vectors of query sketch s_q and the transformed sketches $\widehat{s}_{p_n}(n = 1, 2, \cdots, N)$ are obtained as the input of the cross-domain similarity measurement module.

After performing the above procedure, we can extract the features of the transformed photo $\widehat{p}_{s_q}$, fashion photos $p_n\ (n = 1, 2, \cdots, N)$, the transformed sketches $\widehat{s}_{p_n}(n = 1, 2, \cdots, N)$, and query sketch s_q. Then, we can measure the similarity between the fashion sketch and photo.

3.5. Cross-Domain Similarity Measurement Module. In this section, we measure the similarity of the obtained feature vectors. For the sketch-based fashion photo retrieval stream, the similarity Sim_P between the feature vector $F_{\widehat{p}_{s_q}}$ extracted from the transformed photo $\widehat{p}_{s_q}$ and the feature vectors F_{p_n}

(a)

(b)

(c)

Figure 3: Continued.

(d)

Figure 3: Examples of sketch-based fashion image retrieval results on our newly created dataset by using our proposed model. For each figure, the first column shows the query sketch, and the rest shows the top-10 retrieved fashion photos. The true matches are displayed in the box. (a) Examples of sketch-based clothes retrieval results. (b) Examples of sketch-based pant retrieval results. (c) Examples of sketch-based skirt retrieval results. (d) Examples of sketch-based shoe retrieval results.

extracted from the fashion photos $p_n\,(n = 1, 2, \cdots, N)$ is calculated as

$$\mathrm{Sim}_P = \frac{\sum_{d=1}^{512} F_{\hat{p}_{s_q}}^d \times F_{p_n}^d}{\sqrt{\sum_{d=1}^{512}\left(F_{p\wedge_{s_q}}^d\right)^2} \times \sqrt{\sum_{d=1}^{512}\left(F_{p_n}^d\right)^2}}. \tag{11}$$

For the sketch-based fashion sketch retrieval stream, the similarity Sim_S between the feature vector F_{s_q} extracted from the sketch s_q and the feature vectors $F_{\hat{s}_{p_n}}$ extracted from the transformed sketches $\hat{s}_{p_n}\,(n = 1, 2, \cdots, N)$ is calculated as

$$\mathrm{Sim}_S = \frac{\sum_{d=1}^{512} F_{s_q}^d \times F_{\hat{s}_{p_n}}^d}{\sqrt{\sum_{d=1}^{512}\left(F_{s_q}^d\right)^2} \times \sqrt{\sum_{d=1}^{512}\left(F_{s\wedge_{p_n}}^d\right)^2}}. \tag{12}$$

We assign different weights to balance the influence of these two similarities on the overall similarity and then add them to get the final similarity, which can be expressed as

$$\mathrm{Sim}_{\mathrm{final}} = \mu_1 \mathrm{Sim}_P + \mu_2 \mathrm{Sim}_S, \tag{13}$$

where $\mu_1 = 0.38$ and $\mu_2 = 0.62$ are used in our experiments.

Finally, the relevant fashion photos from the dataset can be returned to the user according to the $\mathrm{Sim}_{\mathrm{final}}$.

4. Experiments and Results

4.1. Experimental Settings

4.1.1. Dataset Preprocessing. There are 12,603 cloth sketch-photo pairs, 5,610 pant sketch-photo pairs, 13,321 skirt sketch-photo pairs, and 4,540 shoe sketch-photo pairs in our introduced Fashion Image dataset. Of these, we use 11,803/4,810/12,321/3,540 pairs for training clothes/pants/skirts/shoes, respectively, and the rest for testing. Before we conduct the experiments, we adjust all the sketches and photos into a unified size of 256×256. In the testing phase, in order to make the sketch in our test set closer to the free-hand sketch, we erased them to remove the details as much as possible, retained the rough outline, and then tested.

Table 2: The retrieval accuracy on our Fashion Image dataset.

Category	acc.@1	acc.@10
Clothes	0.966	0.994
Pants	0.921	0.966
Skirts	0.910	0.971
Shoes	0.905	0.978

We also conduct experiments on two fine-grained instance-level SBIR datasets, i.e., QMUL-shoes and QMUL-chairs datasets [8]. The QMUL-shoes dataset contains 419 shoe sketch-photo pairs, and we use 300 pairs for training and 119 pairs for testing when training our model. The QMUL-chairs dataset contains 297 chair sketch-photo pairs, and we use 200 pairs for training and the rest for testing.

4.1.2. Implementation Details. We used the open source PyTorch to train our models. During training, we use the Adam solver with a batch size of 1. The initial learning rate is set to 0.0001, and momentums are set to 0.5 and 0.999. The maximum number of training iterations is set to 470,000 when training on our Fashion Image dataset. Our method is implemented by NVIDIA Tesla P4 GPU and Intel E5-2630 CPU.

4.1.3. Evaluation Metric. In order to evaluate the performance of our sketch-based fashion image retrieval task, we use retrieval accuracy, denoted as "*acc.@K*." It means the proportion of all the search tasks that can rank the true-match photos in the top K search results.

Table 3: Retrieval accuracy versus different training iterations.

Test	Clothes		Pants		Skirts		Shoes	
Train	Top-1	Top-10	Top-1	Top-10	Top-1	Top-10	Top-1	Top-10
440,000	0.949	0.986	0.920	*0.967*	0.877	0.941	*0.910*	*0.980*
470,000	*0.966*	*0.994*	*0.921*	0.966	*0.910*	*0.971*	0.905	0.978
500,000	0.963	0.991	0.905	0.955	0.874	0.937	0.891	0.974

Figure 4: Retrieval examples obtained by using a complex sketch and a simple sketch for retrieval.

4.2. Experiments on our Fashion Image Dataset. We first conduct retrieval experiments on our Fashion Image dataset for clothes, pants, skirts, and shoes. Figure 3 shows results of our proposed model on the four fashion image retrieval tasks.

4.2.1. Clothes Transformation between Photos and Sketches. We use 11,803 clothes sketch-photo pairs for training the clothes cross-domain transformation model, and the rest for testing. When we obtain the clothes model, we use the model to transform 12,603 Clothes photos to their corresponding clothes sketches, which becomes the transformed clothes sketches dataset.

4.2.2. Pant Transformation between Photos and Sketches. We use 5,610 pant sketch-photo pairs in our Fashion Image dataset to learn to transform pant photos to their corresponding pant sketches. After we get pants transformation model, we use the model to transform 5,610 pant photos to their pant sketches, which forms the transformed pant sketches dataset.

4.2.3. Skirt Transformation between Photos and Sketches. We also use the images of skirts in our Fashion Image dataset to learn to transform skirt images between skirt photos and skirt sketches. After we obtain the skirt transformation, we transform 13,321 skirt photos to 13,321 skirt sketches, which is the transformed skirt sketches dataset.

4.2.4. Shoe Transformation between Photos and Sketches. Finally, we use 3,540 shoe sketch-photo pairs for training the shoe transformation model, and the rest for testing. When we obtain the shoe transformation model, we use the model to transform 4,540 shoe photos to 4,540 shoe sketches, which is the transformed shoe sketches dataset.

After the above experiments, we can get (1) a clothes/pant/skirt/shoe transformation model, respectively, and (2) a transformed fashion sketches dataset consisting of 12,603 transformed clothes sketches, 5,610 transformed pant sketches, 13,321 transformed skirt sketches, and 4,540 transformed shoe sketches.

4.2.5. Sketch-Based Clothes/Pant/Skirt/Shoe Retrieval. Given a clothes/pant/skirt/shoe sketch as a query sketch, firstly, we use the clothes/pant/skirt/shoe transformation model to transform the clothes/pant/skirt/shoe sketch into clothes/pant/skirt/shoe sketch to retrieve the translated fashion sketches dataset, respectively. Therefore, for the query sketch, we perform two retrievals and calculate the weighted sum of the two retrieval results to obtain the final retrieval result based on the clothes/pant/skirt/shoe sketch.

As shown in Table 2, we can find that compared with the correct match in the top-1, the correct match in the top-10 is a much easier task. For sketch-based clothes retrieval, our model ranks the correct match in the top-1 96.6% of the times for clothes. For sketch-based pant retrieval, the pant retrieval accuracy of top-1 and top-10 on our Fashion Image dataset are 92.1% and 96.6%. For sketch-based skirt retrieval, the top-1 and top-10 retrieval accuracy are up to 91.0% and 97.1%. As for sketch-based shoe retrieval, the accuracy of the true-match shoe photo ranked in the top-1 and top-10 are 90.5% and 97.8%. Figure 3 shows several retrieval results of our proposed model on our contributed dataset, the left part of the figure shows the query sketches, and the right part shows the top-10 retrieved fashion photos. If there are true-match photos in the top-10, most of their positions are in the top-1.

Finally, we used different types of fashion images to conduct experiments and analysed the impact of the training iterations to the retrieval accuracy. For different training iterations on the cross-domain fashion image transformation experiments, we calculated the retrieval accuracy achieved in different training iterations. The results were reported in Table 3. We found that, when the training iterations in the cross-domain fashion image transformation experiment phase were 470,000 iterations, the overall performance of

Table 4: Accuracy comparison with baselines on the Fashion Image dataset, QMUL-shoes and QMUL-chairs datasets.

Methods	Fashion Image dataset		QMUL-shoes		QMUL-chairs	
	acc.@1	acc.@10	acc.@1	acc.@10	acc.@1	acc.@10
BoW-HOG + rank-SVM [6]	—	—	0.174	0.678	0.289	0.670
ISN Deep + rank-SVM [37]	—	—	0.200	0.626	0.474	0.825
Dense-HOG + rank-SVM [8]	—	—	0.244	0.652	—	—
3DS Deep + rank-SVM [10]	—	—	0.052	0.217	0.062	0.268
Sketchy [17]	0.673	0.953	—	—	—	—
Our model	0.924	0.977	0.308	0.654	0.495	0.794

Figure 5: Examples of query sketch retrieval results on QMUL-shoes and QMUL-chairs datasets by using our proposed model.

retrieval accuracy in the test phase is the best, i.e., the top-1 retrieval accuracy for clothes, pants, skirts, and shoes is 96.6%, 92.1%, 91.0%, and 90.5%, respectively. What is more, our Fashion Image dataset contains sketches of different styles and different complexity, and some sketches have problems such as noise, unclear images, and missing strokes. We used these sketches to test and found that no matter how complex or simple the input sketch of the model is, the model can achieve good retrieval performance. Some retrieval results are shown in Figure 4.

4.3. Comparison with Baselines. We conduct experiments with baselines on three datasets: our Fashion Image dataset, QMUL-shoes, and QMUL-chair datasets [8]. The baselines we selected include Sketchy [17], BoW-HOG + rank-SVM [6], Improved Sketch-a-Net (ISN) [37], Dense-HOG + Rank-SVM [8], and 3D shape (3DS) [10]. Compared with baselines, our model transforms the sketches and photos to the same domain before retrieval, which improves the retrieval accuracy to a certain extent. The detailed comparative experiment results are shown in Table 4.

4.3.1. Comparison with Baselines on Our Fashion Image Dataset. We compare our model with Sketchy on our newly created Fashion Image dataset. Table 4 shows the top-1 and top-10 retrieval accuracy comparison with baseline on our

Table 5: Impact of sketch-based fashion photo retrieval stream and sketch-based fashion sketch retrieval stream on retrieval accuracy, implemented on our Fashion Image, QMUL-shoes, and QMUL-chairs datasets.

Methods	Fashion Image dataset		QMUL-shoes		QMUL-chairs	
	acc.@1	acc.@10	acc.@1	acc.@10	acc.@1	acc.@10
(1) Sketch-based fashion photo retrieval stream only	0.612	0.811	0.192	0.462	0.381	0.660
(2) Sketch-based fashion sketch retrieval stream only	0.911	0.957	0.154	0.500	0.351	0.680
(3) Our full model	0.924	0.977	0.308	0.654	0.495	0.794

dataset. As it shows in this table, our approach outperforms the Sketchy by 25.1% and 2.4% in top-1 retrieval accuracy and top-10 retrieval accuracy, respectively.

4.3.2. Comparison with Baselines on the QMUL-Shoes Dataset. In addition to experimentally comparing our approach with the baselines on our newly created dataset, we also evaluate our approach on the QMUL-shoes dataset. The QMUL-shoes dataset is a fine-grained instance-level SBIR dataset which contains 419 shoe sketch-photo pairs. On this dataset, we compare our model with BoW-HOG + rank-SVM, Improved Sketch-a-Net (ISN), Dense-HOG + Rank-SVM, and 3D shape (3DS). We compare our method with baselines in terms of top-1 and top-10 accuracies on QMUL-shoes dataset. From Table 4, we can find that our model can achieve compelling performance on the QMUL-shoes dataset and outperform the Dense-HOG + rank-SVM by 6.4% in top-1 retrieval accuracy. Examples of query sketch retrieval results on QMUL-shoes are presented in Figure 5.

4.3.3. Comparison with Baselines on the QMUL-Chairs Dataset. The QMUL-chairs dataset contains 297 chair sketch-photo pairs. We also conduct experiments on this fine-grained instance-level SBIR dataset. We compare our model with BoW-HOG + rank-SVM, Improved Sketch-a-Net (ISN), and 3D shape. In Table 4, we present the top-1 and top-10 accuracies of our model over other three models on the QMUL-chairs dataset for fine-grained SBIR. Compared with other methods, the top-1 retrieval accuracy of our model is higher than ISN Deep + rank-SVM by 2.1%. Examples of the query sketch and top-10 retrieval results on QMUL-chairs dataset are shown in Figure 5.

4.4. Ablation Studies. In this section, in order to demonstrate the advantage of combining the sketch-based fashion photo retrieval stream with the sketch-based fashion sketch retrieval stream, we conduct three ablation studies on our Fashion Image, QMUL-shoes, and QMUL-chairs datasets. Table 5 shows the results. The three ablation studies are as follows: (1) Only the sketch-based fashion photo retrieval stream is used, and the sketch-based fashion sketch retrieval stream is not used for retrieval. From Table 5, we find that on our Fashion Image dataset, the top-1 retrieval accuracy is 61.2%, and the top-10 retrieval accuracy is 81.1%. (2) Instead of using the sketch-based fashion photo retrieval stream, only the sketch-based fashion sketch retrieval stream is used for retrieval. As shown in Table 5, on our Fashion Image dataset, the retrieval accuracy of the top-1 is 91.1% and that of the top 10 is 95.7%. (3) Our full model of combining the two methods is combines the sketch-based fashion photo retrieval stream and the sketch-based fashion sketch retrieval stream for the ablation study. As shown in Table 5, the retrieval accuracy of top-1 reaches the highest on all three datasets, i.e., 92.4%, 30.8%, and 49.5%, respectively. After the above ablation studies, from Table 5, we can draw the conclusion that combining the two retrieval streams has further improved the retrieval results.

5. Conclusions and Future Work

In this paper, we first contributed a Fashion Image dataset, which contains 36,074 sketch-photo pairs for conducting research on sketch-based fashion image retrieval. We then introduced a new algorithm for sketch-based fashion image retrieval based on cross-domain transformation, which improves the retrieval accuracy by fusing the sketch-based fashion photo retrieval stream and sketch-based fashion sketch retrieval stream. Among them, the sketch-based fashion photo retrieval stream is to transform the query sketch into the corresponding photo in the natural photo domain and then use the transformed photo to retrieve the dataset. The sketch-based fashion sketch retrieval stream is to transform the fashion photos in the dataset to the corresponding sketches in the sketch domain and then use the query sketch to retrieve the transformed sketch dataset. The two similarities obtained by these two methods are first weighted, then added to obtain a hybrid similarity, and finally use the hybrid similarity for sketch-based fashion image retrieval.

Also, the current network has limitation that some sketches cannot be transformed into ideal photos. In future work, we will collect more fashion images of different styles and commit ourselves to research a network that can transform simple sketches into ideal photos and improve retrieval accuracy.

Data Availability

The Fashion Image data used to support the findings of this study are available from the corresponding author upon request, and the QMUL-shoes and the QMUL-chairs data used to support the findings of this study are available from this website: http://www.eecs.qmul.ac.uk/~qian/Project_cvpr16.html.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Acknowledgments

This research was jointly supported by the National Natural Science Foundation of China (61762050, 61876074, 61877031) and China Scholarship Council (201908360112). The authors would like to thank Fan Yang from the School of Computer and Information Engineering, Jiangxi Normal University, for his help in experimental design.

References

[1] Y. Kalantidis, L. Kennedy, and L. J. Li, "Getting the look: clothing recognition and segmentation for automatic product suggestions in everyday photos," in *Proceedings of the 3rd ACM Conference on International Conference on Multimedia Retrieval*, pp. 105–112, Dallas, USA, 2013.

[2] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang, "Deepfashion: powering robust clothes recognition and retrieval with rich annotations," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1096–1104, Las Vegas, NV, USA, June 2016.

[3] T. Xiao, T. Xia, Y. Yang, C. Huang, and X. Wang, "Learning from massive noisy labeled data for image classification," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2691–2699, Boston, MA, USA, June 2015.

[4] M. H. Kiapour, X. Han, S. Lazebnik, A. C. Berg, and T. L. Berg, "Where to buy it: matching street clothing photos in online shops," in *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 3343–3351, Santiago, Chile, December 2015.

[5] X. Liang, L. Lin, W. Yang, P. Luo, J. Huang, and S. Yan, "Clothes co-parsing via joint image segmentation and labeling with application to clothing retrieval," *IEEE Transactions on Multimedia*, vol. 18, no. 6, pp. 1175–1186, 2016.

[6] Y. Li, T. M. Hospedales, Y. Z. Song, and S. Gong, "Free-hand sketch recognition by multi-kernel feature learning," *Computer Vision and Image Understanding*, vol. 137, pp. 1–11, 2015.

[7] M. Eitz, J. Hays, and M. Alexa, "How do humans sketch objects?," *ACM Transactions on Graphics*, vol. 31, no. 4, pp. 1–10, 2012.

[8] Q. Yu, F. Liu, Y. Z. Song, T. Xiang, T. M. Hospedales, and C. C. Loy, "Sketch me that shoe," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 799–807, Las Vegas, NV, USA, June 2016.

[9] K. Pang, K. Li, Y. Yang et al., "Generalising fine-grained sketch-based image retrieval," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 677–686, Long Beach, CA, USA, June 2019.

[10] F. Wang, L. Kang, and Y. Li, "Sketch-based 3d shape retrieval using convolutional neural networks," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1875–1883, Boston, MA, USA, June 2015.

[11] L. Liu, F. Shen, Y. Shen, X. Liu, and L. Shao, "Deep sketch hashing: fast free-hand sketch-based image retrieval," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2862–2871, Honolulu, Hawaii, USA, July 2017.

[12] R. Hu and J. Collomosse, "A performance evaluation of gradient field hog descriptor for sketch based image retrieval," *Computer Vision and Image Understanding*, vol. 117, no. 7, pp. 790–806, 2013.

[13] J. M. Saavedra and B. Bustos, "An improved histogram of edge local orientations for sketch-based image retrieval," in *Pattern Recognition. DAGM 2010. Lecture Notes in Computer Science, vol 6376*, M. Goesele, S. Roth, A. Kuijper, B. Schiele, and K. Schindler, Eds., pp. 432–441, Springer, Berlin, Heidelberg, 2010.

[14] J. Collomosse, T. Bui, and H. Jin, "Livesketch: query perturbations for guided sketch-based visual search," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2879–2887, Long Beach, CA, USA, June 2019.

[15] Y. Li, T. M. Hospedales, Y. Z. Song, and S. Gong, "Fine-grained sketch-based image retrieval by matching deformable part models," *British Machine Vision Association, BMVA*, 2014.

[16] P. Xu, Q. Yin, Y. Huang et al., "Cross-modal subspace learning for fine-grained sketch-based image retrieval," *Neurocomputing*, vol. 278, pp. 75–86, 2017.

[17] P. Sangkloy, N. Burnell, C. Ham, and J. Hays, "The sketchy database: learning to retrieve badly drawn bunnies," *ACM Transactions on Graphics*, vol. 35, no. 4, p. 119, 2016.

[18] K. Li, K. Pang, Y. Z. Song, T. M. Hospedales, T. Xiang, and H. Zhang, "Synergistic instance-level subspace alignment for fine-grained sketch-based image retrieval," *IEEE Transactions on Image Processing*, vol. 26, no. 12, pp. 5908–5921, 2017.

[19] S. Zou, W. Chen, and H. Chen, "Image classification model based on deep learning in internet of things," *Wireless Communications and Mobile Computing*, vol. 2020, Article ID 6677907, 16 pages, 2020.

[20] S. Chen, M. Wang, and X. Chen, "Image annotation via reconstitution graph learning model," *Wireless Communications and Mobile Computing*, vol. 2020, Article ID 8818616, 9 pages, 2020.

[21] Ş. Öztürk, "Image inpainting based compact hash code learning using modified U-Net," in *2020 4th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*, pp. 1–5, Istanbul, Turkey, October 2020.

[22] C. Cui, X. Wu, J. Yang, and J. Li, "A novel DIBR 3D image hashing scheme based on pixel grouping and NMF," *Wireless Communications and Mobile Computing*, vol. 2020, Article ID 8820436, 14 pages, 2020.

[23] Ş. Öztürk, "Two-stage sequential losses based automatic hash code generation using Siamese network," *European Journal of Science and Technology*, vol. 1, pp. 39–46, 2020.

[24] M. A. Shah, N. Y. Khanday, M. Purohit, and M. H. Gulzar, *Enhancement and Segmentation of Lung CT Images for Efficient Identification of Cancerous Cells*, 2016.

[25] Ş. Öztürk, "Stacked auto-encoder based tagging with deep features for content-based medical image retrieval," *Expert Systems with Applications*, vol. 161, p. 113693, 2020.

[26] N. Y. Khanday and S. A. Sofi, "Taxonomy, state-of-the-art, challenges and applications of visual understanding: a review," *Computer Science Review*, vol. 40, article 100374, 2021.

[27] Q. Yu, Y. Yang, Y. Z. Song, T. Xiang, and T. Hospedales, "Sketch-a-Net that beats humans," in *British Machine Vision Conference 2015*, pp. 1–12, Swansea, UK, 2015.

[28] Q. Chen, J. Huang, R. Feris, L. M. Brown, J. Dong, and S. Yan, "Deep domain adaptation for describing people based on fine-grained clothing attributes," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5315–5324, Boston, MA, USA, June 2015.

[29] I. Goodfellow, J. Pouget-Abadie, M. Mirza et al., "Generative adversarial nets," *In Proceedings of the 27th International Conference on Neural Information Processing Systems*, pp. 2672–2680, 2014.

[30] J. Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2223–2232, Venice, Italy, October 2017.

[31] M. Y. Liu, T. Breuel, and J. Kautz, "Unsupervised image-to-image translation networks," 2017, https://arxiv.org/abs/1703.00848.

[32] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, http://arxiv.org/abs/1409.1556.

[33] P. Dollar and C. L. Zitnick, "Fast edge detection using structured forests," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 8, pp. 1558–1570, 2014.

[34] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," 2013, http://arxiv.org/abs/1312.6114.

[35] D. J. Rezende, S. Mohamed, and D. Wierstra, "Stochastic backpropagation and variational inference in deep latent gaussian models," *International Conference on Machine Learning*, vol. 2, 2014.

[36] A. B. L. Larsen, S. K. Sønderby, H. Larochelle, and O. Winther, "Autoencoding beyond pixels using a learned similarity metric," *Proceedings of Machine Learning Research*, vol. 48, pp. 1558–1566, 2016.

[37] Q. Yu, Y. Yang, F. Liu, Y. Z. Song, T. Xiang, and T. M. Hospedales, "Sketch-a-Net: a deep neural network that beats humans," *International Journal of Computer Vision*, vol. 122, no. 3, pp. 411–425, 2017.