Neurocomputing 341 (2019) 156–167

Contents lists available at ScienceDirect

Neurocomputing

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

Toward AI fashion design: An Attribute-GAN model for clothing match

Linlin Liu [a], Haijun Zhang [a,*], Yuzhu Ji [a], Q.M. Jonathan Wu [b]

[a] Department of Computer Science, Harbin Institute of Technology, Shenzhen, China
[b] Department of Electrical and Computer Engineering, University of Windsor, Ontario, Canada

ARTICLE INFO

Article history:
Received 1 November 2018
Revised 19 February 2019
Accepted 7 March 2019
Available online 12 March 2019

Communicated by Dr Zhao Zhang

Keywords:
Clothing match
Generative adversarial network
Fashion data
Attribute

ABSTRACT

Dressing in clothes based on the matching rules of color, texture, shape, etc., can have a major impact on perception, including making people appear taller or thinner, as well as exhibiting personal style. Unlike the extant fashion mining literature, in which style is usually classified according to similarity, this paper investigates clothing match rules based on semantic attributes according to the generative adversarial network (GAN) model. Specifically, we propose an Attribute-GAN to generate clothing-match pairs automatically. The core of Attribute-GAN constitutes training a generator, supervised by an adversarial trained collocation discriminator and attribute discriminator. To implement the Attributed-GAN, we built a large-scale outfit dataset by ourselves and annotated clothing attributes manually. Extensive experimental results confirm the effectiveness of our proposed method in comparison to several state-of-the-art methods.

1. Introduction

Fashion is a form of expression that can highlight each person's personality when he or she is authentic with individual style choices. Coco Chanel expressed this succinctly when she stated, "Fashion fades, only style remains the same." Traditional fashion design relies on the designer's personal creative sense, which can possess uncertainty and subjectivity. With the advent of the Big Data era, fashion design patterns have changed. Indeed, fashion style can now be analyzed by machine learning from images and textual descriptions of clothing, shoes, or accessories. Fashion trends are the result of numerous factors, such as color, shape, texture, patterns, etc. The items in an outfit should also be at least subtly compatible regarding these factors. Fig. 1 presents several attractive outfits that are composed of a set of clothes. Although they are all fine collocations, these outfits have very different styles.

Observing people's dressing habits, we find that whether or not an outfit matches is mostly determined by clothing attributes, e.g., boxy matches stretchy; light blue matches white; panels match stripes, etc. It is possible to mine these rules of clothing match through artificial intelligence (AI). Specifically, this paper aims to elucidate latent match rules considering clothing attributes under the framework of the generative adversarial network (GAN). These latent match rules are then utilized to generate outfit composition.

Fashion learning has recently attracted great attention in the computer vision field due to its vast lucrative applications. In fact, a large body of literature exists that focuses on clothing segmentations [10,14,34,36,37,39], recognition [15,17], and fashion image retrieval [7,8,20–22]. Instead of assigning a semantic clothing label to each pixel of a person in an image, some other works have focused on identifying fashion ability [28,38] or occupation [30] from the clothing in images. In addition, some researchers have explored methods for clothing retrieval, including within-scenario retrieval [20] and cross-scenario retrieval [21,22]. However, modeling fashion collocation presents certain challenges. On the one hand, fashion style is subtle and subjective. Consequently, since the sensitivity of fashion varies from person to person, it can be difficult to unify the labeled data. On the other hand, it is quite challenging to obtain a detailed and complete set of attributes to describe whether or not a match is stylish. As a result, few existing studies focus on identifying why one outfit looks good, and then provide advice for creating a well composed outfit. Since Goodfellow et al. [9] proposed the generative adversarial network (GAN) in 2014, various derivative GAN models have been proposed. The innovative components of these models include model structure improvement [6,24,25,27,35,41], theoretical extension [1–3,16,40,42], and applications [12,32,35,43,44]. In this paper, we pilot the use of artificial intelligence (AI) in the fashion industry by developing an Attribute-GAN model. We propose Attribute-GAN to generate clothes, which takes a fashion outfit associated with clothing attributes as the input. The generator is trained to produce clothing pairs. However, two adversarial trained discriminators are respectively used who

* Corresponding author.
E-mail address: hjzhang@hit.edu.cn (H. Zhang).

https://doi.org/10.1016/j.neucom.2019.03.011
0925-2312/

Fig. 1. Attractive outfits which have very different styles.

can predict whether or not the clothing pairs match or whether or not the attributes of fake clothing in these pairs are correct. The trained GAN strengthens the power of the generator to generate pairs that match on some attributes. This is more consistent with people's dressing habits and aesthetics, instead of style similarity.

To evaluate the effectiveness of our proposed model, we built an outfit dataset containing over 160,000 clothing images, approximately 40,000 of which were annotated with attributes by human labelers. The attributes were compiled from key terms frequently retrieved in several major e-commerce sites. For evaluation, we employed two tactics, a subjective method and an objective method, for evaluating the authenticity of fake images and the collocation degree of generated clothing pairs. The subjective method constituted conducting a "real or fake" study on human labelers; whereas, the objective method included using a regression model to score the matching degree of the generated outfit and training an inception model to calculate the inception score of fake images similar to [27]. Extensive experimental results demonstrate the effectiveness of our method in comparison to state-of-the-art models.

2. Related work

There are many bodies of related work; we focus on attributes learning, compatibility learning and GAN model.

2.1. Attributes learning

As described in Section 1, compatibility between fashion items usually relies on semantic attributes. Attribute learning has been widely investigated in numerous computer vision studies. Related to the fashion domain, attribute learning has been utilized for image retrieval [11,22], fine-grained categorization [5], and sentence generation from clothing [4]. In [11], researchers proposed a dual attribute-aware ranking network (DARN) to integrate semantic attributes with visual similarity constraints, and then to model the discrepancy between images from different domains. Liu et al. [22] also presented a new deep model, FashionNet, to predict clothing attributes by using a weighted cross-entropy loss, and performed in-shop clothes retrieval and consumer-to-shop clothes retrieval by employing a triplet loss function. They defined their deep model according to specific tasks and used the image as a whole for feature extraction. Some other works in attribute learning extracted features based on a pre-detected bounding box, such as [5], in which a regional-convolutional neural network (R-CNN) framework was utilized to detect human bodies, and then a specific double-path deep neural network was proposed by modeling the two domains with separate paths. Moreover, Chen et al. [4] extracted several types of features with respect to human poses. Instead of employing deep models, however, they used support vector machines (SVMs) to learn attributes, and utilized conditional random fields to capture mutual dependencies between two attributes.

2.2. Compatibility learning

Some extant literature has also evaluated compatibility between fashion items [13,19,33]. Specifically, Li et al. proposed a multi-model deep learning framework, which uses context information and clothing images jointly to evaluate outfit quality [19]. Iwata et al. detected top and bottom regions by using pose estimation, and then utilized a topic model to learn co-occurrence information about visual features in top and bottom regions [13]. Veit et al. trained a Siamese CNN to learn style space and to generate outfits based on the learned model [33]. These methods adopted different models to evaluate the compatibility of different categories of images. However, they all only considered identifying some common characteristics or style similarities of items in outfits. In this research, we assert that good clothing collocation may be in line with peoples' aesthetics and habits. In other words, it does not necessarily possess common characteristics. Therefore, our aim is to elucidate latent collocation rules based on clothing attributes, and then generate matching items from given clothes.

2.2.1. GAN model

The generative adversarial network (GAN) [9] was inspired from the zero-sum game in game theory. In a zero-sum game, the sum of the interests of two players is zero or one constant, i.e., one party gains and the other loses. The two players in the GAN model are a generative model and a discriminative model, respectively. The generative model captures the distribution of data samples; whereas, the discriminative model is a classifier, which estimates the possibility that a sample belongs to the training data. The generator model and discriminative model generally constitute nonlinear mapping functions, such as multi-layer perceptions and

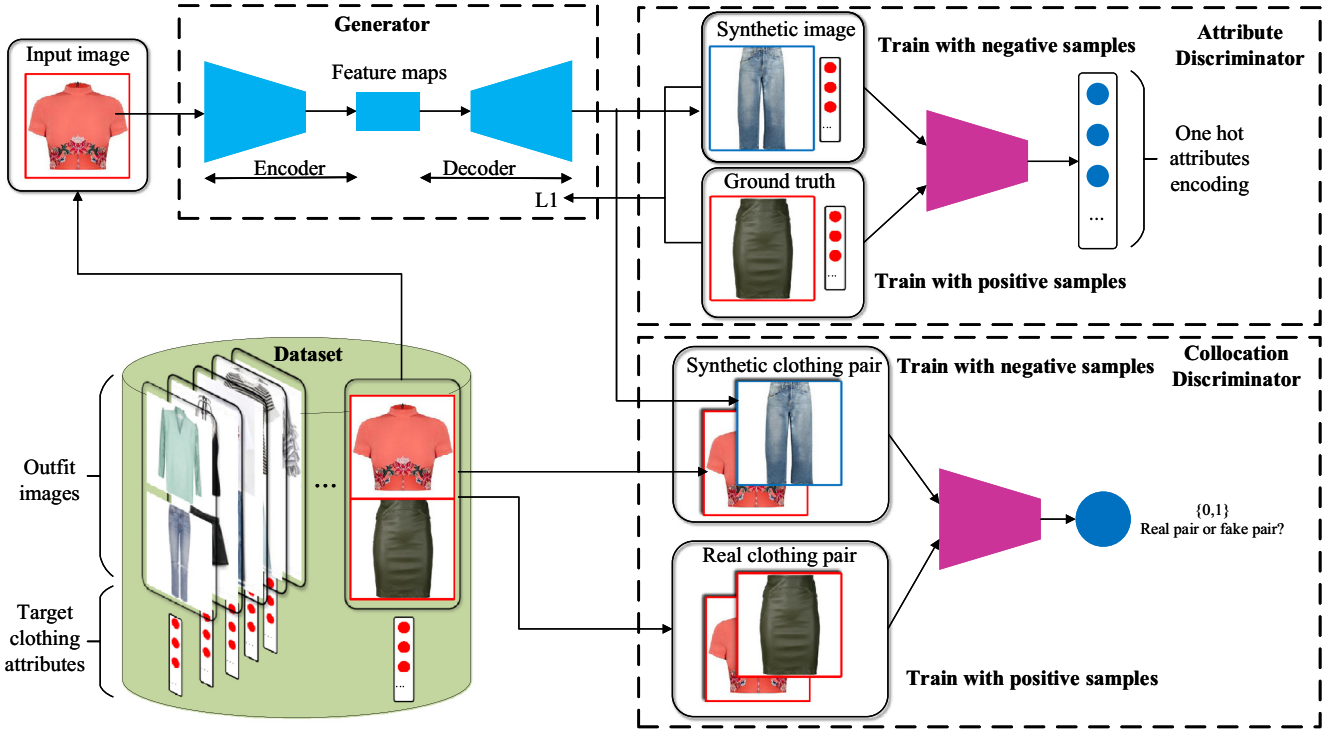

Fig. 2. An overview of the Attribute-GAN model (The output of attribute discriminator is a vector encoded by one hot attribute values predicted by attribute discriminator. The collocation discriminator determines whether or not the fake clothing matches the input real image. In the output of collocation discriminator, "1" denotes that input image pairs are real and "0" denotes that input image pairs are fake.).

CNNs. In recent years, a large body of research has been proposed to improve the theoretical basis of GANs. For example, several works have attempted to add certain explicit external information as partial guidance to solve the issue of freedom in GAN training process [12,23,27,35,43]. Especially, conditional GAN [23] (cGAN) makes the generator and discriminator conditioned on some additional information. Some investigations have also relied on fine-grained guidance by partitioning the generation process into many steps [6,24,25,41]. Laplacian generative adversarial networks (LAPGANs) [6] constitute the first effort to apply the hierarchical generating process to GAN. Another problem of GANs is the vanishing of gradient due to a drawback of the object function. Several studies have attempted to overcome this issue [1,16,40,44]. One representative work is [1], in which Wasserstein distance was utilized to improve learning stability.

3. Attribute-GAN model

Our model can be regarded as an extension of cGAN, which learns a mapping from label x and random noise vector z, to y: G: $z \rightarrow y$. Our proposed approach, Attribute-GAN, learns a mapping from a pair of outfits, conditioned on attributes of clothing. Both the generator network G and the discriminator network D perform feed-forward inference conditioned on clothing attributes. As shown in Fig. 2, the model consists of three components, i.e., a generator and two discriminators. Inspired by Isola et al. [12], we employed a popular convolutional encoder-decoder architecture, U-Net [26], as a generator. There are two discriminators. One of them is a convolutional 'PatchGAN' discriminator, which is able to capture high-frequency structural information in local patches [12,18]. The other discriminator is a multi-task attribute classifier network, which determines whether or not a generated fake clothing image has the expected ground truth attributes.

3.1. Objective function

We propose a variant of the cGAN architecture, termed Attribute-GAN. The whole architecture of Attribute-GAN is illustrated in Fig. 2. The generator network can be represented as $G : \mathbb{R}^X \rightarrow \mathbb{R}^Y$, while the collocation discriminator is denoted as $D_{\text{collo}} : \mathbb{R}^X \times \mathbb{R}^Y \rightarrow \{0, 1\}$; the attribute discriminator is denoted as $D_{\text{attri}} : \mathbb{R}^Y \rightarrow \{1, ..., M\}$, where X is the dimension of the given clothing, Y is the dimension of the output matched clothing, and M is the number of attributes.

In cGAN, both generator and discriminator are conditioned on certain additional information y, which could be any kind of auxiliary information. Here, in our model, y is a collocation of clothing. The discriminator in our model is responsible for predicting whether or not a clothing pair matches, instead of classifying a generated image as real or fake in traditional cGAN. During the training process in our model, we view image pairs as joint observations to train the collocation discriminator. The objective function under a conditional GAN framework can be defined in the form of:

$$L_{\text{collo}}(G, D_{\text{collo}}) = E_{x,y\sim P_{\text{data}}(x,y)}[\log D_{\text{collo}}(x, y)] + E_{x\sim P_{\text{data}}(x)}[\log(1 - D_{\text{collo}}(x, G(x)))], \quad (1)$$

where P_{data} is generator's distribution.

In order to fool the discriminator, we keep the discriminator unchanged, but add an $L1$ distance to the generator objective [12]. Therefore, a generated image should be nearer to the ground truth clothing considering the impact of $L1$ distance. In the final objective function (see Eq. (4)), the loss associated with the $L1$ distance

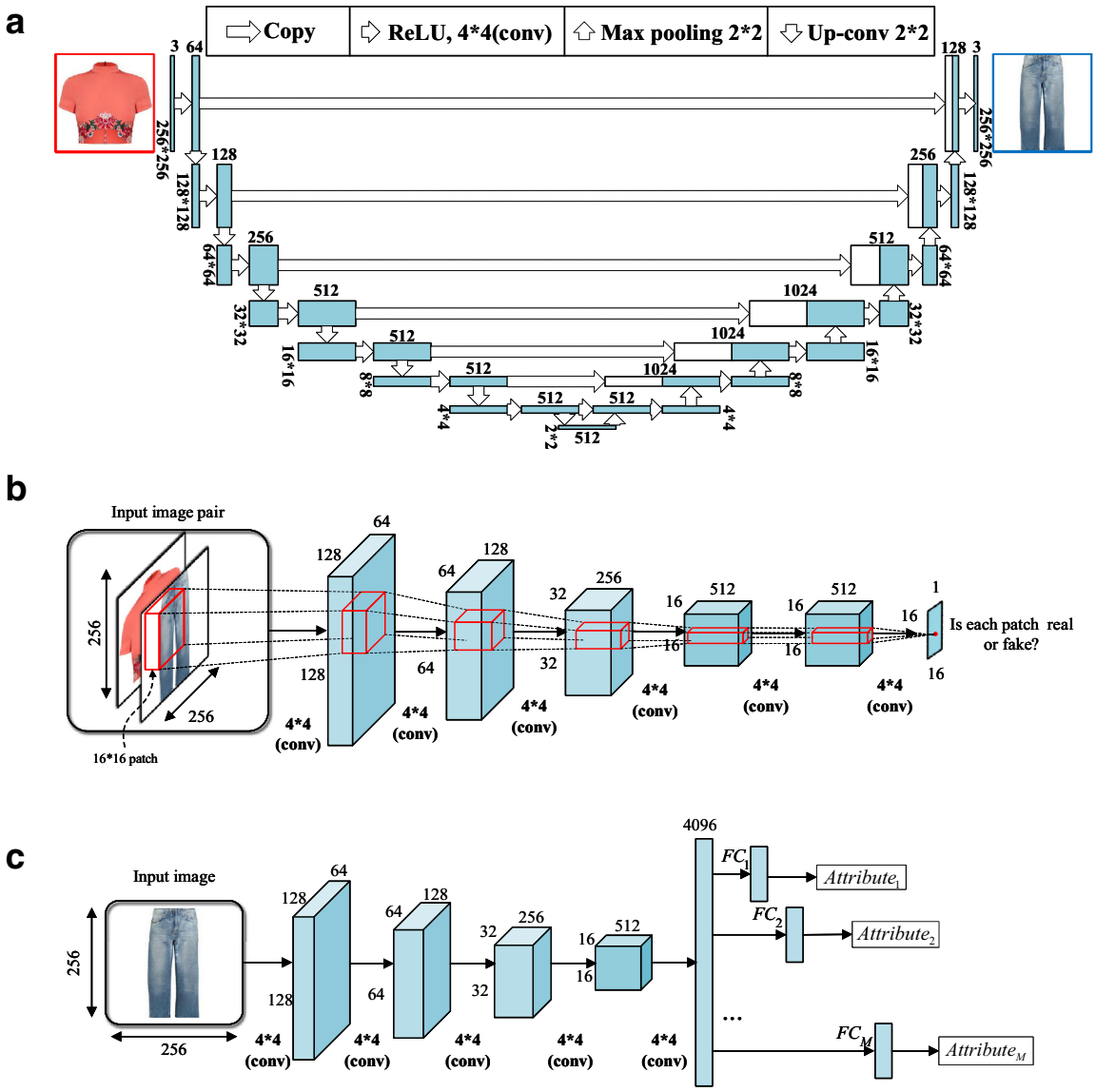

Fig. 3. The detailed network structure of: (a) generator; (b) collocation discriminator; and (c) attribute discriminator.

will be multiplied by a hyper-parameter, λ, to balance all the loss terms. The objective function for the generator becomes:

$$L_{L1}(G) = E_{x,y\sim p_{\text{data}}(x,y)}[\|y - G(x)\|_1]. \tag{2}$$

Targeting to generate a suitable fashion outfit, the generated clothing is tasked to be close to the attributes of ground truth collocation clothing. So, we design another attribute-hitting discriminator to produce a probability distribution over the attributes. By learning to optimize the attribute-hitting process, the discriminator can provide an additional signal to the generator. The attribute discriminator framework is presented in Fig. 3. It is worth noting that this framework can be considered for application to several multi-class classification tasks. The objective function for the attribute discriminator framework can be computed by:

$$L_{\text{attri}}(G, D_{\text{attri}}) = \frac{1}{2M}\sum_{i=1}^{M} E[\log P(A_i = a_i|y)] + E\left[\log P(A_i = a_i|\hat{y})\right], \tag{3}$$

where A_i denotes the ith clothing attribute; and a_i denotes the ith attribute ground truth.

Thus, the final objective function of our proposed Attribute-GAN is:

$$G^* = \arg\min_G \max_D L_{\text{collo}}(G, D_{\text{collo}}) + \lambda L_{L1}(G) + L_{\text{attri}}(G, D_{\text{attri}}) \tag{4}$$

Under this extension framework to cGAN, the training process is supervised by the collocation images pairs and semantic clothing attributes. It aims to strictly learn a map between an image and a matched image visually as well as the latent matching rules of attributes.

3.2. Inference and optimization

The entire training procedure is summarized in Algorithm 1. Instead of generating a fake image from random noise, we generate the fake image $\hat{y}$ (shown in line 3) from a real clothing image. After forwarding through the discriminator by real image pairs and fake image pairs, we obtain their matching degree score, i.e., s_r and s_f (shown in lines 4 and 5). For restricting the generator in the attributes latent style space, we add another attribute discriminator to indicate whether the fake image has the attributes that

Algorithm 1 Attribute-GAN training algorithm.

1: **Input**: mini batch collocation image pairs (x, y), attributes of clothing A_y, number of training batch steps S, step size α
2: **Output**: Attribute-GAN generator model
3: **for** n = 1 to S **do**
4: $\widehat{y} \leftarrow G(x)$ forward through generator
5: $s_r \leftarrow D_{\text{collo}}(x, y)$ forward through collocation discriminator by real clothing image pairs
6: $s_f \leftarrow D_{\text{collo}}(x, \widehat{y})$ forward through collocation discriminator by fake clothing image pairs
7: $p_r \leftarrow D_{\text{attri}}(y, A_y)$ forward through attribute discriminator by real clothing images and their attributes
8: $p_f \leftarrow D_{\text{attri}}(\widehat{y}, A_y)$ forward through attribute discriminator by fake clothing images and real attributes
9: $L_{D_{\text{collo}}} \leftarrow \log(s_r) + \log(1 - s_f)$
10: $L_{D_{\text{attri}}} \leftarrow \frac{1}{2M} \sum_{j=1}^{M} (CE(p_{rj}) + CE(p_{fj}))$
11: $D_{\text{collo}} \leftarrow D_{\text{collo}} - \alpha \partial L_{D_{\text{collo}}} / \partial D_{\text{collo}}$ update collocation discriminator
12: $D_{\text{attri}} \leftarrow \alpha \partial L_{D_{\text{attri}}} / \partial D_{\text{attri}}$ update attribute discriminator
13: $L_G \leftarrow \log(s_f) + \frac{1}{M} \sum_{j=1}^{M} (CE(p_{fj})) + \|y - \widehat{y}\|_1$
14: $G \leftarrow G - \alpha \partial L_G / \partial G$ update generator
15: **end for**

the corresponding real collocation input clothing tends to have. Thus, we get p_r and p_f (shown in lines 6 and 7). *CE* (shown in lines 9 and 12) denotes the cross entropy loss. Note that the generator's tasks include fooling the collocation discriminator, fooling the attribute discriminator, and making generated clothing close to the real collocation image in the $L1$ sense. Lines 11, 12 and 14 in Algorithm 1 indicate taking a gradient step to update network parameters.

We employed a standard approach from to optimizing the network [9], that is, alternating one gradient descent step on collocation discriminator, one step on attribute discriminator, and then one step on generator. Mini batch SGD (stochastic gradient descent) and Adam solver were adopted in above training process.

4. Experiments

4.1. Dataset

In order to train and evaluate our proposed model, Attribute-GAN, we compiled a large-scale dataset from Ployvore,[1] which is a free, easy-to-use web-based application for mixing and matching images from anywhere on the Internet. Users create fashion outfits with items on the website or item images uploaded by themselves. The items in a fashion outfit collaged according to users' preferences aim to beautifully exhibit specific fashion styles. In this platform, tonality tags or text tags added by the users can be searched, scored, shared, commended, and recreated by a visitor. We crawled fashion outfits and their related information from the website www.ployvore.com. For each outfit, we collected the images of items, title, category, number of likes, etc. In this research, only images and the number of likes were utilized.

In this work, after annotation and manual data cleaning, we obtained 19,081 pairs of collocation clothing images, including upper clothes and lower clothes with their attributes. 15,000 pairs were selected for model training, 3000 pairs were used for validation, and 1081 pairs were utilized for testing. In addition to the labeled pairs, we crawled an additional 81,103 image pairs with counts of visitors who like the outfits. Although these images were not annotated with attributes manually, they were utilized for further model evaluation.

[1] https://www.polyvore.com/.

4.2. Attribute annotation

After investigating frequently searched indexes in several major e-commerce websites, we established a set of fine-grained attributes of clothing, including category, color, texture, shape, etc. However, the original images collected from www.ployvore.com did not contain attribute information. In order to train and evaluate our model, ten graduate students who are majored in computer science were invited to manually annotate the attributes of items in order to describe clothing from a variety of views as compactly and comprehensively as possible. Nine types of clothing attributes were extracted, and the total number of attributes was 93. The attributes list in Table 1. And the attribute distribution in the training dataset is shown in Fig. 4. As shown in Fig. 5, each image is labeled by attributes, such as tile texture, pattern, type of model, form, etc.

4.3. Parameter settings and compared models

For the implementation details of our network structure, we adopted U-net with skip connections as a generator and a 'PatchGAN' classifier as a collocation discriminator under a pix2pixGAN [12] structure. For the attribute discriminator, we designed a five-layer CNN to classify the attributes of clothing images. At the final layer, a full connection layer was applied to map the number of output channels, i.e., the number of each attribute. In addition, the sum of the attribute losses is back-propagated. The detailed architectures are presented in Fig. 3. In the training stage, input samples were firstly resized to 286×286, and were then cropped to 256×256. Specifically, we set the number of epochs to 200. The batch size was set to 1. The learning rate was initialized to 0.0002, and an Adam solver with momentum 0.5 was used. The hyperparameter λ was set to 100, it will be discussed in the Section 4.6.

In the Attribute-GAN model, the task is generating the matching image pairs based on the compatibility rules of a training set. In order to exhibit the effectiveness of Attribute-GAN, we compare our result with the state-of-the-art methods, including cGAN+$L1$ [12], cGAN [23], Vanilla GAN [9], the only generator trained by $L1$ loss. These classical GAN models with attribute discriminator are designed to test and verify the effectiveness of attribute discriminator. Specifically, the compared methods are cGAN, Vanilla GAN, cGAN+$L1$, Vanilla GAN+$L1$, cGAN+D_{attri}, Vanilla GAN+D_{attri}, Vanilla GAN+$L1$+D_{attri}. The objective function under a Vanilla GAN framework can be defined in the form of:

$$L_{\text{GAN}}(G, D_{\text{vanilla}}) = E_{y \sim P_{\text{data}}(y)}[\log D_{\text{vanilla}}(y)] + E_{x \sim P_{\text{data}}(x)}[1 - \log D_{\text{vanilla}}(G(x))]. \quad (5)$$

Again, similar with Attribute-GAN, the final objective function of compared Vanilla GAN+$L1$+D_{attri} is defined in the form of:

$$G^* = \arg \min_G \max_D L_{\text{GAN}}(G, D_{\text{vanilla}}) + \lambda L_{L1}(G) + L_{\text{attri}}(G, D_{\text{attri}}). \quad (6)$$

To fully evaluate the capability of GAN models for such a task, firstly we performed two types of experiment: lower clothing generation given upper clothing, and upper clothing generation conditioning on lower clothing. Representative examples are illustrated in Fig. 6.

We can observe that the generator trained only with $L1$ loss usually leads to fuzzy clothing shape images. Training with cGAN based model without $L1$ loss in generator, however, will generate the same pattern, which is the skirt shape combined with each

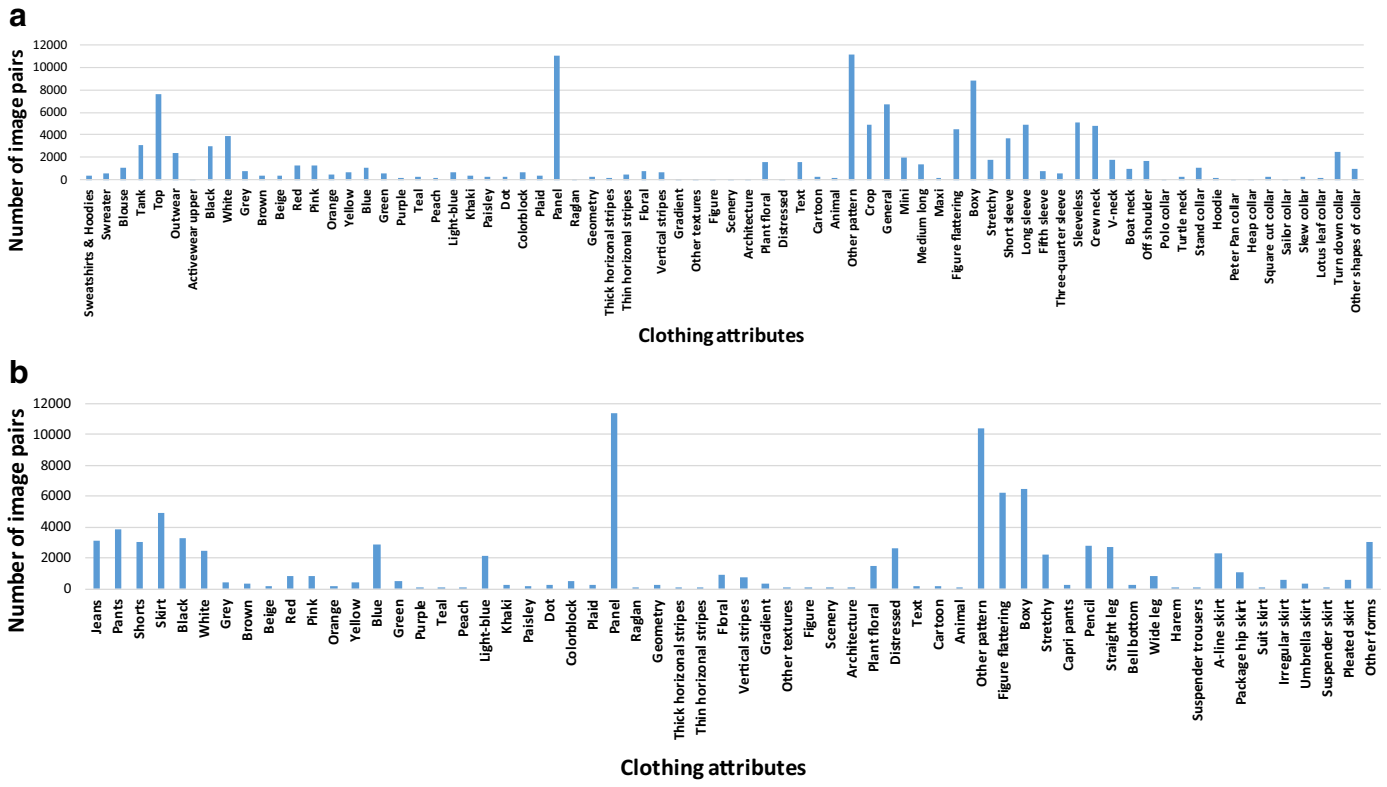

Fig. 4. Attribute distribution over the training dataset with respect to: (a) upper clothing; (b) lower clothing.

Fig. 5. Semantic attributes for clothing.

Table 1
Clothing attribute statistics (there are 93 attributes in total for clothing, including 9 classes, each of which contains multi-class attributes).

Attribute Name	Attribute value	
	Upper clothing	Lower clothing
Category	Sweatshirts and Hoodies, Sweater, Blouse, Tank, Tops, Outwears, Activewear upper	Jeans, Pants, Shorts, Skirts
Color	Black, white, gray, brown, beige, red, pink, orange, yellow, blue, green, purple, teal, peach, light blue, khaki	
Texture	Paisley, dotted, color block, plaid, panel, geometric, thick horizontal stripes, thin horizontal stripes, floral, vertical stripes, gradient, other textures	
Pattern	Figures, numbers, scenery, architecture, plant floral, distressed, text, cartoon, animal, other patterns	
Mode	Figure flattering, boxy, stretchy	
Form	–	Capri pants, pencil, straight leg, bell bottom, wide leg, harem, suspender trousers, A-line skirt, package hip skirt, suit skirt, irregular skirt, umbrella skirt, suspender skirt, pleated skirt, other forms
Size	Crop, general, mimi, medium long, maxi	–
Shape of sleeve	Short sleeve, long sleeve, fifth sleeve, Three quarter sleeve, sleeveless	–
Shape of collar	Crew neck, V neck, boat neck, off shoulder, polo collar, turtle neck, stand collar, hoodie, peter pan collar, heap collar, square cut collar, sailor collar, skew collar, lotus leaf collar, turn down collar, other shape of collar	–

corresponding upper clothing outline, or meaningless images, i.e., it suffers from serious mode omission. Observing the cGAN loss values, the discriminator loss tends to be zero at last, which indicates that the discriminator is too strong. As a result, it always predicts that the synthetic pairs are fake and the ground truth pairs are real. Considering the situation without an $L1$ constraint, since there is always a real image in an image pair, the discriminator more easily identifies a fake image pair than a single image, when the discriminator predicts a pair of images instead of single image in Vanilla GAN.

As shown in Fig. 6, for lower clothing generation conditioning on upper clothing, the Attribute-GAN and compared models can produce results that are able to mix the fake with the genuine. In particular, for generating clothing images with simple mode and style, our model performs quite well so that the generated clothing could be very similar to real images. In addition, we observe that the result of Attribute-GAN is superior in terms of categorical diversity, style diversity, and clothing fineness. We also observe that the quality of generated jeans and pants are better than that of generated dress. We believe this phenomenon is ascribed to the complex patterns and rich diversities of dress. To overcome this limitation, it would be worth building a larger dataset with balanced attribute distributions in order to improve the generalization ability of our model in future work. Considering that upper clothes may have various styles and delicate details, the synthetic upper clothing images are not generated well in comparison to the synthetic lower clothing. According to our observations, the task of upper clothing generation given lower clothing is still very

Fig. 6. Examples produced by compared methods.

Table 2
Inception scores for generating images by different models.

	Mean	Std
Ground truth	1.238	0.0085
Attribute-GAN	**1.230**	**0.0097**
cGAN+$L1$	1.207	0.0120
Vanilla GAN	1.114	0.0110
Vanilla GAN+$L1$	1.204	0.0100
Vanilla GAN+D_{attri}	1.114	0.0102
Vanilla GAN+$L1$+D_{attri}	1.195	0.0142

challenging for all the models. Therefore, in the following context we only demonstrated performance evaluation on the task with respect to lower clothing generation given upper clothing. In the following section, we employed two aspects, authenticity and compatibility, to fully evaluate the performance of Attribute-GAN.

4.4. Results on authenticity

Firstly, we measure the authenticity of generated images as the basic goal of GAN. However, distinguishing between realistic images and artificial ones is incredibly difficult. We perform this in two ways, an objective method and a subjective method. In the objective way, similar to [27], we trained an inception-v3 model [31] to calculate inception scores in order to quantify the quality and diversity of generated images. In the subjective way, we conducted a user study to ask several annotators to distinguish generated data from real data.

4.4.1. Qualitative results

We adopted an automatic method by calculating the "inception score" for quantitative evaluation: $\mathrm{I} = \exp\left(E_x D_{KL}\left(p(y|x)\|p(y)\right)\right)$, where x denotes the generated image, and y denotes the label predicted by the inception-v3 model. Since the dataset that we used does not have an existing pre-trained inception model, we trained one based on the dataset that we built. In the dataset we crawled, each item includes an image and its category. We selected the categories from the dataset based on the attributes for annotating. Here, we only measure the visual quality of synthetic lower clothing. Thus, we only used the categories of lower clothes by the inception-v3 model, which have four classes. For each class, we randomly selected 10,000 images. In the test phase, the result of our pre-trained model with respect to precision reaches to 93%.

Table 2 quantifies the image quality using Inception score. Here, we compared the algorithms which are able to generate effective images. It is observed that Attribute-GAN achieves the highest scores, which are close to that of the ground truth. It indicates that the attribute discriminator in the objective function shown in Eq. (4) encourages the output to respect the input on image details, which makes the generated image look more realistic. We can observe that the synthetic images generated by cGAN+$L1$-based objective have more details, for example, pocket, distressed hole, fold, etc. as shown in Fig. 6.

4.4.2. User study

In subjective way, we conducted a user study by building a Web interface as illustrated in Fig. 7. Five users (not including any of the authors) were asked to participate in the experiment by evaluating eight images at each time. These images are the ground truth collocation lower clothing images and fake images generated by our algorithm and other compared methods. The annotators should choose which one is fake.

The results are listed in Table 3, in which *FP* indicates the number of synthetic images which are labeled as real clothing images, and *FN* means the number of real clothing images that are regarded as fake clothing images. The results in Table 3 means the proportion of mislabeling synthetic clothing images in all mistakes that human annotators made. Our model has higher result which means the synthetic clothing images generated by Attribute-GAN have higher authenticity and have stronger ability to fool the human labelers.

Table 3
Incorrect rate of distinguishing the synthetic images.

	FP/*FP*+*FN*
Attribute-GAN	**0.654**
cGAN+$L1$	0.610
Vanilla GAN	0.651
Vanilla GAN+$L1$	0.484
Vanilla GAN+D_{attri}	0.508
Vanilla GAN+$L1$+D_{attri}	0.445

Table 4
Statistics results produced by regression model.

	MAE	$S_f \geq S_r$
Attribute-GAN	**0.097**	**918/1081**
cGAN+$L1$	0.100	906/1081
Vanilla GAN	0.104	894/1081
Vanilla GAN+$L1$	0.105	895/1081
Vanilla GAN+D_{attri}	0.104	914/1081
Vanilla GAN+$L1$+D_{attri}	0.104	895/1081

In addition, Fig. 8 shows the category distribution statistics of generating images. We can observe that Attribute-GAN exhibits encouraging results, because the categories covered by the clothing are well-balanced categories distributed. Under the condition on same category distribution in a training set, Attribute-GAN performs well with respect to the diversity generalization under the supervision of attribute discriminator.

4.5. Results on compatibility

In this section, we evaluate the matching degree of generated pairs. The same as evaluating the quality of synthesized images, we employed both an objective and a subjective method. The objective method trains a regression model to score the generated clothing image pairs. The subjective method performs a user study by asking human annotators to score the matching degree.

4.5.1. Qualitative results

For an objective evaluation, we trained a regression CNN model to predict the matching degree in terms of generated image pairs. Here, we adjusted the VGG-16 model [29] to fit with the scoring task. Concretely, we assigned the dimension of the output of the final full connection layer to be one dimension, and used the mean square error (MSE) loss to update the network. Given the like count of a fashion outfit, we regarded the like count as the score of compatibility, and input to the regression model as a kind of supervision information. The numbers of training image pairs and testing pairs were 80,000 and 1103, respectively. After 50 epochs, the loss value was no longer descending. Then, we evaluated the regression model with MAE (Mean Absolute Error) setting at 0.15 on the testing set.

The target of our model is generating collocation pairs which have the same matching degree as ground truth in terms of matching score. In Table 4, the first column represents the MAE values between the scores of ground truth pairs and those of generated pairs. We can observe that Attribute-GAN outperforms other methods. In the second column, S_f denotes the synthetic

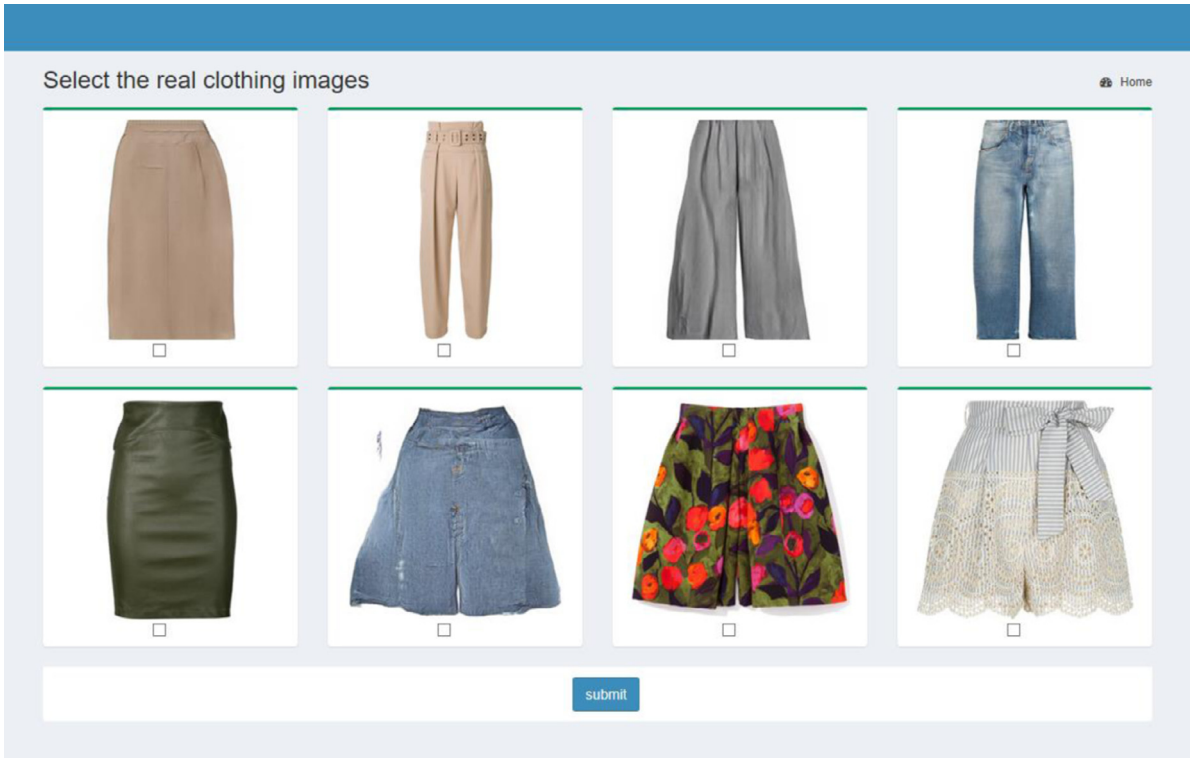

Fig. 7. Web interface given to annotators who were asked to distinguish synthetic clothing images among real ones.

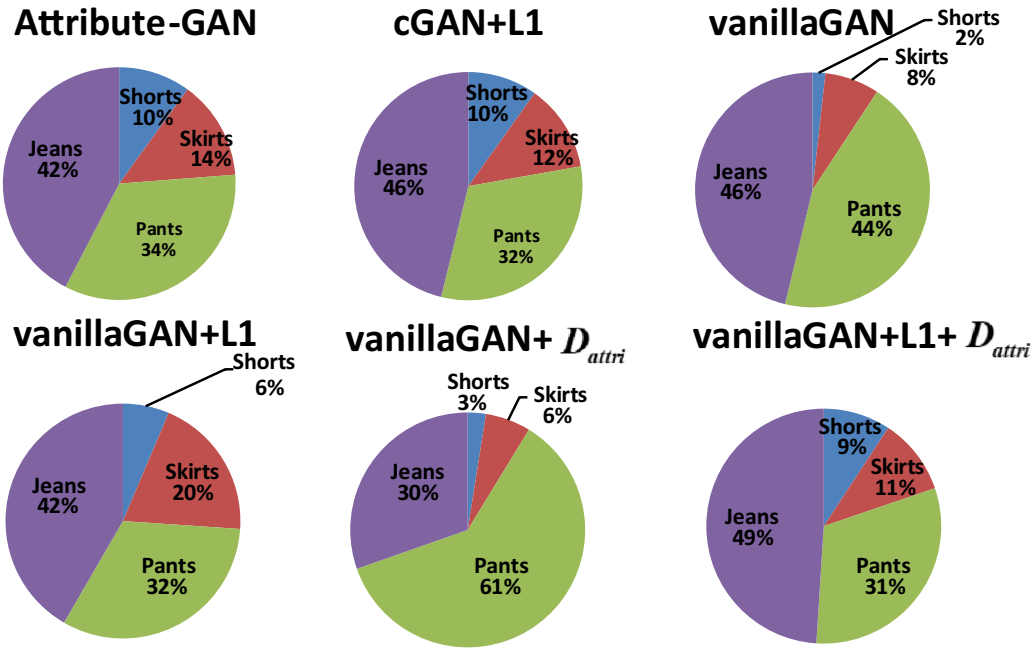

Fig. 8. Category distribution of generated images using different models.

pair score, while S_r denotes the score coming from ground truth pairs. The results based on this measure indicate that the number of synthetic clothing pairs whose scores are predicted by regression model is higher than that of the corresponding ground truth clothing pairs. With the supervision provided by attributes discriminator, the synthetic clothing pairs from Attribute-GAN have a superior matching degree in comparison to the ground truth.

4.5.2. User study

For a user study on the compatibility of generated pairs, we utilized the same subjects employed in Section 4.5. For each fake pair, we set five degrees from one to five to score compatibility. Each labeler should choose one number to represent the matching level. The used website for this evaluation is shown in Fig. 9.

The evaluation results are summarized in Table 5, in which degree 5 means the best compatibility degree of clothing outfit, while

Fig. 9. Web interface given to subjects, who should choose the compatibility degree from one to five, and should label the category of lower clothing.

Table 5
Average rate of compatibility of real clothing pairs and synthetic clothing pairs in user study (degree 1 is the worst and 5 is the best).

	1	2	3	4	5
Attribute-GAN	0.103	0.243	0.349	0.198	0.108
cGAN+$L1$	0.147	0.212	0.275	0.286	0.081
Vanilla GAN	0.077	0.167	0.379	0.255	**0.123**
Vanilla GAN+$L1$	0.170	0.254	0.254	0.295	0.027
Vanilla GAN+D_{attri}	0.108	0.225	0.308	0.345	0.014
Vanilla GAN+$L1$+D_{attri}	0.146	0.215	0.263	0.315	0.062

1 indicates the worst. From the overall performance, our method, Attribute-GAN, produces better results over cGAN+$L1$ in terms of the compatibility of generated clothing pairs. Our model, however, does not show advantage over vanilla GAN, because vanilla GAN has the largest rate at degree 5. From Fig. 8, under the condition on same category distributions in a training set, we can observe that Attribute-GAN performs the best on diversity, while vanilla GAN performs the worst, as 90% of generated clothing of vanilla GAN are pants and jeans. Without considering conditional discriminator and attribute discriminator, vanilla GAN has mode missing problem in the training process. The simple generated type will make the result of user study deceptive. For example, jeans are more easily to match different types of top clothing. Thus, nonprofessional users will give higher scores to the outfits with jeans. So it is reasonable that vanilla GAN would perform better to some.

4.6. Parametric study

From Fig. 6, we can see that $L1$ loss seems to be a very important component in the cGAN-based models. Without $L1$ loss, cGAN-based models will suffer from serious mode omission. As we discussed earlier, the $L1$ constraint guides the generator in a right direction to transform the input images. Then the collocation discriminator will identify whether synthetic clothing pairs are good or not, and the attribute discriminator gives the prediction on semantic attributes. Here, we will verify the significance of $L1$ loss by experimenting on different settings of the hyper-parameter λ in the Attribute-GAN model. Fig. 10 gives a few examples of qualitative results. It is observed that Attribute-GAN with λ that is larger than 50 will generate effective clothing images. But with the increase of this parameter, the performance may degrade. Quantitative inception scores reported in Table 6 confirm our conjecture that a large value of $L1$ loss may significantly weaken the capability of generator to deliver collocation information and attribute semantics hidden in clothing images pairs. In real-world applications, setting the parameter, λ, to 100 is sufficient to produce satisfactory results for our Attribute-GAN model.

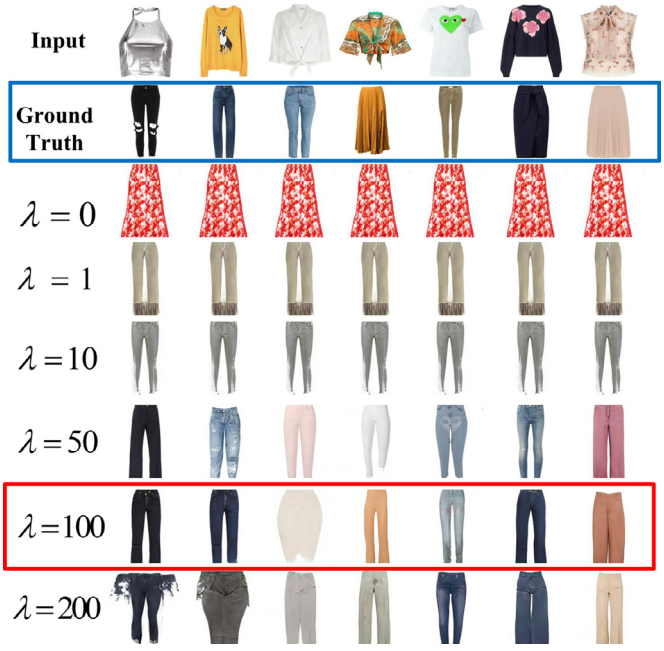

Fig. 10. Qualitative examples of hyper-parameter λ discussion.

Table 6
Inception scores of the effective results produced by Attribute-GAN model with different hyper-parameter λ.

	$\lambda = 50$	$\lambda = 100$	$\lambda = 200$
Mean	1.204	**1.226**	1.199
Std	0.0114	**0.0066**	0.0102

5. Conclusion

This paper investigates the clothing match problem under the cGAN framework. Specifically, we proposed an Attribute-GAN model, a scalable image-to-image translation model between different domains by a generator and two discriminators, which generate collocation clothing images based on semantic attributes. Besides the advantage of generating higher image quality, Attribute-GAN achieved the best diversity of synthetic images and matching degree of generated clothing outfits, owing to the generalization capability behind the supervision of attribute discriminator and collocation discriminator. In addition, we built an attributes labeled outfits dataset to evaluate the effectiveness of our model. In future work, we plan to augment attributes dataset, and extend more applications to clothing retrieval and recommendations.

Conflicts of interest

No conflict of interest exits in the submission of this manuscript, and manuscript is approved by all authors for publication. I would like to declare on behalf of my co-authors that the work described was original research that has not been published previously, and not under consideration for publication elsewhere, in whole or in part. All the authors listed have approved the manuscript that is enclosed.

Acknowledgments

This work was supported in part by the National Key R&D Program of China under grant no. 2018YFB1003800, 2018YFB1003805, the Natural Science Foundation of China under grant no. 61832004, and the Shenzhen Science and Technology Program under grant no. JCYJ20170413105929681.

References

[1] M. Arjovsky, S. Chintala, L. Bottou, in: Wasserstein gan, 2017. preprint arXiv: 170107875.
[2] D. Berthelot, T. Schumm, L. Metz, in: Began: boundary equilibrium generative adversarial networks, 2017. preprint arXiv: 170310717.
[3] T. Che, Y. Li, A.P. Jacob, Y. Bengio, W. Li, in: Mode regularized generative adversarial networks, 2016. preprint arXiv: 161202136.
[4] H. Chen, A. Gallagher, B. Girod, Describing clothing by semantic attributes, in: European Conference on Computer Vision, 2012, pp. 609–623.
[5] Q. Chen, J. Huang, R. Feris, L.M. Brown, J. Dong, S. Yan, Deep domain adaptation for describing people based on fine-grained clothing attributes, in: IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 5315–5324.
[6] E.L. Denton, S. Chintala, R. Fergus, et al., Deep generative image models using a Laplacian pyramid of adversarial networks, in: Advances in Neural Information Processing systems, 2015, pp. 1486–1494.
[7] W. Di, C. Wah, A. Bhardwaj, R. Piramuthu, N. Sundaresan, Style finder: fine–grained clothing style detection and retrieval, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2013, pp. 8–13.
[8] J. Fu, J. Wang, Z. Li, M. Xu, H. Lu, Efficient clothing retrieval with semantic-preserving visual phrases, in: Asian Conference on Computer Vision, Springer, 2012, pp. 420–431.
[9] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: Advances in Neural Information Processing Systems, 2014, pp. 2672–2680.
[10] B. Hasan, D.C. Hogg, Segmentation using deformable spatial priors with application to clothing, in: BMVC, 2010, pp. 1–11.
[11] J. Huang, R.S. Feris, Q. Chen, S. Yan, Cross-domain image retrieval with a dual attribute-aware ranking network, in: IEEE International Conference on Computer Vision, 2015, pp. 1062–1070.
[12] P. Isola, J.Y. Zhu, T. Zhou, A.A. Efros, in: Image-to-image translation with conditional adversarial networks, 2016. preprint arXiv: 161107004.
[13] T. Iwata, S. Wanatabe, H. Sawada, Fashion coordinates recommender system using photographs from fashion magazines, in: IJCAI, 22, 2011, p. 2262.
[14] N. Jammalamadaka, A. Minocha, D. Singh, C. Jawahar, Parsing clothes in unrestricted images, in: BMVC, 1, 2013, p. 2.
[15] M.H. Kiapour, K. Yamaguchi, A.C. Berg, T.L. Berg, Hipster wars: discovering elements of fashion styles, in: European Conference on Computer Vision, Springer, 2014, pp. 472–488.
[16] T. Kim, M. Cha, H. Kim, J. Lee, J. Kim, in: Learning to discover cross-domain relations with generative adversarial networks, 2017. preprint arXiv: 170305192.
[17] I.S. Kwak, A.C. Murillo, P.N. Belhumeur, D.J. Kriegman, S.J. Belongie, From bikers to surfers: visual recognition of urban tribes, in: BMVC, 1, 2013, p. 2.
[18] C. Li, M. Wand, Precomputed real-time texture synthesis with Markovian generative adversarial networks, in: European Conference on Computer Vision, Springer, 2016, pp. 702–716.
[19] Y. Li, L. Cao, J. Zhu, J. Luo, Mining fashion outfit composition using an end–to-end deep learning approach on set data, in: IEEE Transactions on Multimedia, 2017.
[20] S. Liu, J. Feng, Z. Song, T. Zhang, H. Lu, C. Xu, S. Yan, Hi, magic closet, tell me what to wear!, in: Proceedings of the 20th ACM International Conference on Multimedia, ACM, 2012a, pp. 619–628.
[21] S. Liu, Z. Song, G. Liu, C. Xu, H. Lu, S. Yan, Street-to-shop: cross-scenario clothing retrieval via parts alignment and auxiliary set, in: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2012b, pp. 3330–3337.
[22] Z. Liu, P. Luo, S. Qiu, X. Wang, X. Tang, Deepfashion: powering robust clothes recognition and retrieval with rich annotations, in: IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1096–1104.
[23] M. Mirza, S. Osindero, in: Conditional generative adversarial nets, 2014. preprint arXiv:14111784.
[24] A. Nguyen, J. Yosinski, Y. Bengio, A. Dosovitskiy, J. Clune, in: Plug & play generative networks: conditional iterative generation of images in latent space, 2016. preprint arXiv: 161200005.
[25] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, H. Lee, in: Generative adversarial text to image synthesis, 2016. preprint arXiv: 160505396.
[26] O. Ronneberger, P. Fischer, T. Brox, U-net: convolutional networks for biomedical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2015, pp. 234–241.
[27] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, X. Chen, Improved techniques for training gans, in: Advances in Neural Information Processing Systems, 2016, pp. 2234–2242.
[28] E. Simo-Serra, S. Fidler, F. Moreno-Noguer, R. Urtasun, Neuroaesthetics in fashion: modeling the perception of fashionability, in: IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 869–877.
[29] K. Simonyan, A. Zisserman, in: Very deep convolutional networks for large-scale image recognition, 2014. preprint arXiv: 14091556.
[30] Z. Song, M. Wang, X.S. Hua, S. Yan, Predicting occupation via human clothing and contexts, in: IEEE International Conference on Computer Vision, IEEE, 2011, pp. 1084–1091.
[31] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2818–2826.
[32] Y. Taigman, A. Polyak, L. Wolf, in: Unsupervised cross-domain image generation, 2016. preprint arXiv: 161102200.
[33] A. Veit, B. Kovacs, S. Bell, J. McAuley, K. Bala, S. Belongie, Learning visual clothing style with heterogeneous dyadic co-occurrences, in: IEEE International Conference on Computer Vision, 2015, pp. 4642–4650.
[34] N. Wang, H. Ai, Who blocks who: Simultaneous clothing segmentation for grouping images, in: IEEE International Conference on Computer Vision, IEEE, 2011, pp. 1535–1542.
[35] H. Wu, S. Zheng, J. Zhang, K. Huang, in: Gp-gan: towards realistic high-resolution image blending, 2017. preprint arXiv: 170307195.
[36] K. Yamaguchi, M.H. Kiapour, L.E. Ortiz, T.L. Berg, Parsing clothing in fashion photographs, in: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2012, pp. 3570–3577.
[37] K. Yamaguchi, M. Hadi Kiapour, T.L. Berg, Paper doll parsing: retrieving similar styles to parse clothing items, in: IEEE International Conference on Computer Vision, 2013, pp. 3519–3526.
[38] K. Yamaguchi, T.L. Berg, L.E. Ortiz, Chic or social: visual popularity analysis in online fashion networks, in: Proceedings of the 22nd ACM international conference on Multimedia, ACM, 2014, pp. 773–776.
[39] W. Yang, P. Luo, L. Lin, Clothing co-parsing by joint image segmentation and labeling, in: IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 3182–3189.
[40] Z. Yi, H. Zhang, P.T. Gong, et al., in: Dualgan: unsupervised dual learning for image-to-image translation, 2017. preprint arXiv: 170402510.
[41] H. Zhang, T. Xu, H. Li, S. Zhang, X. Huang, X. Wang, D. Metaxas, in: Stackgan: text to photo-realistic image synthesis with stacked generative adversarial networks, 2016. preprint arXiv: 161203242.
[42] J. Zhao, M. Mathieu, Y. LeCun, in: Energy-based generative adversarial network, 2016. preprint arXiv: 160903126.
[43] J.Y. Zhu, P. Krähenbühl, E. Shechtman, A.A. Efros, Generative visual manipulation on the natural image manifold, in: European Conference on Computer Vision, Springer, 2016, pp. 597–613.
[44] J.Y. Zhu, T. Park, P. Isola, A.A. Efros, in: Unpaired image-to-image translation using cycle-consistent adversarial networks, 2017. preprint arXiv: 170310593.

Linlin Liu received the B.S. degree in computer science from Zhengzhou University of Aeronautics, Zhengzhou, China, in 2012, and the M.S. degree in computer science from the Harbin Institute of Technology, Shenzhen, China, in 2016, where she is currently pursuing the Ph.D. degree in computer science. Her research interests include data mining, computer vision, image processing, and deep learning.

Haijun Zhang received the B.Eng. and Master's degrees from Northeastern University, Shenyang, China, and the Ph.D. degree from the Department of Electronic Engineering, City University of Hong Kong, Hong Kong, in 2004, 2007, and 2010, respectively. He was a Post-Doctoral Research Fellow with the Department of Electrical and Computer Engineering, University of Windsor, Windsor, ON, Canada, from 2010 to 2011. Since 2012, he has been with the Harbin Institute of Technology, Shenzhen, China, where he is currently an associate professor of computer science. His current research interests include multimedia data mining, machine learning, and computational advertising. He is currently an associate editor of neurocomputing, neural computing and applications, and pattern analysis and applications.

Yuzhu Ji received the B.S. degree in computer science from PLA Information Engineering University, Zhengzhou, China, in 2012, and the M.S. degree in computer engineering from the Harbin Institute of Technology, Shenzhen, China, in 2015, where he is currently pursuing the Ph.D. degree in computer science. His research interests include data mining, computer vision, image processing, and deep learning.

Q.M. Jonathan Wu received the Ph.D. degree in electrical engineering from the University of Wales, Swansea, U.K., in 1990. He was with the National Research Council of Canada for ten years from 1995, where he became a senior research officer and a group leader. He is currently a professor with the Department of Electrical and Computer Engineering, University of Windsor, Windsor, ON, Canada. He has published more than 250 peer-reviewed papers in computer vision, image processing, intelligent systems, robotics, and integrated microsystems. His current research interests include 3-D computer vision, active video object tracking and extraction, interactive multimedia, sensor analysis and fusion, and visual sensor networks. Dr. Wu holds the Tier 1 Canada Research Chair in Automotive Sensors and Information Systems. He was an associate editor of the IEEE Transactions onSystems, Man, and Cybernetics Part A and the International Journal of Robotics and Automation. He has served on technical program committees and international advisory committees for many prestigious conferences.