IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 15, 2020, accepted July 25, 2020, date of publication August 3, 2020, date of current version August 14, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3013631

ClothingNet: Cross-Domain Clothing Retrieval With Feature Fusion and Quadruplet Loss

YONGWEI MIAO[1,2], (Member, IEEE), GAOYI LI[3], CHEN BAO[3], JIAJING ZHANG[2], AND JINRONG WANG[1]

[1]College of Information Science and Engineering, Hangzhou Normal University, Hangzhou 311121, China
[2]College of Information Science and Technology, Zhejiang Sci-Tech University, Hangzhou 310018, China
[3]College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou 310023, China

Corresponding author: Yongwei Miao (ywmiao2009@hotmail.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61972458, and in part by the Science Foundation of Zhejiang Sci-Tech University under Grant 17032001-Y.

ABSTRACT Cross-domain clothing retrieval is an active research topic because of its massive potential applications in fashion industry. Due to the large number of garment categories or styles, and different clothing appearances caused by different camera angles, different shooting conditions, different messy background environments, or different postures of the dressed human body, the retrieval accuracy of traditional consumer-to-shop scheme is always low. In this paper, based on the framework of deep convolution neural network, a novel cross-domain clothing retrieval method is proposed by using feature fusion and quadruplet loss function, which is named as ClothingNet. First, the pre-trained deep neural network Resnet-50 is adopted to extract feature map of clothing images. The extracted high-level image features can thus be merged with middle-level features, and the final feature representation of clothing images can be obtained by constraining the fusion feature values to a certain range in term of L_2 norm. This fusion feature provides a comprehensive description of the differences between clothing images. For effectively training our ClothingNet, the cross-domain clothing images are organized in form of a quadruplet for calculating its loss function, and the network parameters can be optimized according to back propagation scheme via stochastic gradient descent of loss function. Our proposed method is validated on two public datasets for clothing retrieval, DARN and DeepFashion, showing that the top-50 retrieval accuracy is 35.67% and 53.52% respectively. Experimental results illustrate the effectiveness of our clothing retrieval method.

INDEX TERMS Clothing retrieval, neural network, feature extraction, feature fusion, quadruplet loss.

I. INTRODUCTION

Nowadays, online clothes shopping has become increasingly popular as a fashion shopping manner for young people. Given a consumer-captured photo of a special garment (from the user consumer domain), the technique of cross-domain clothing retrieval aims to search commercial garment images (from the commercial shop domain) that are corresponding to the detected clothing item [1]. Unlike traditional schemes of image retrieval, the cross-domain retrieval of clothing images requires some similarity measurements of two underlying images that may be captured from different heterogeneous domains, which is always face some difficulties and challenges mainly due to the following reasons: 1) large intra-class clothes variance and also 2) minor inter-class clothes variance. For example, even the same garment category can lead to different styles of clothing due to subtle patterns discrepancy. All of these issues make it difficult to retrieve clothing images across the user consumer domain and the commercial shop domain.

The concept of cross-domain clothing retrieval was first introduced by Liu *et al.* [1]. They firstly obtained the local features of clothing images by extracting 30 regions of human body according to the human pose estimation, which can reduce the image differences due to different dressing postures of cross-domain clothing images. And thus they employed a local feature matching method for clothes retrieval. In order to reduce the influence of complex and messy background environments, Kalantidis *et al.* [2] proposed a method of clothing recognition and regional representation, which adopted a binary space mask for constraining human body areas obtained by the pose estimation algorithm.

The associate editor coordinating the review of this manuscript and approving it for publication was Jeon Gwanggil.

In recent years, with the rapid development of convolution neural network (CNN), the traditional feature representation and feature extraction methods of clothing images have been gradually replaced by the neural network schemes [3], which can learn a general model and can also map the clothing images to special feature values in latent feature space. The feature values that represent the same cross-domain garment image pairs can be as close as possible. Owing to the locations of image key-points, a FashionNet network proposed by Liu *et al*. [4] carried out the cross-domain retrieval of clothing images by using the registered key-points and image attribute information. However, this method spends a lot of labors and also a lot of time to mark the key-points of clothing images. Gajic and Baldrich [5] employed the deep convolution neural network as a feature extractor of clothing images, and also adopted two neural networks with different parameters for sensing their differences between consumer domain and shop domain clothing images. However, if two cross-domain networks adopt completely different network parameters, the number of parameters will be large, which is not conducive to the training and optimization of the underlying neural network. Xiong *et al*. [6] introduced a CNN architecture for effectively retrieving clothing images by capturing their proposed image features of two domains, which can be obtained by partial-sharing network parameters between different domains. In order to reduce the influence of messy background for image retrieval, Ji *et al*. [7] introduced a attention model to focus the training of network on the clothing itself and thus effectively neglect the influence of background noise. Luo *et al*. [8] presented a deep multi-task cross-domain hashing, in which the cross-domain embedding and sequential attribute learning can be modeled simultaneously. Kucer and Murray [9] proposed a detection and then retrieval model for multi-domain fashion item retrieval. Park *et al*. [10] introduced a combined training strategy and loss function optimization method to improve the accuracy of fashion image retrieval.

However, the feature vectors of clothing images in the above retrieval approaches are represented by the final convolution layer output of neural network. With the increase of number of layers in neural network, the intuitive low-level texture information of clothing images will be lost, and the abstract high-level image semantic information will be retained [11]. In our task of cross-domain clothing retrieval, it is necessary to capture both high-level semantic information and also middle-level feature information of clothing images for determining whether two garments match or not. Here, the garment category can be judged by semantic information of clothing images, and the middle-level features of clothing images can be employed to determine the garment style, the garment pattern and the decorative patterns, and so on. Therefore, the image retrieval schemes using high-level semantic features can only suitable for image classification [12] or image recognition tasks [13], which will not suitable for our cross-domain clothing retrieval task. Moreover, for constructing the loss function of neural network, the triplet loss function [14] is adopted in references [4]–[6]. Here, the triplets can be represented as the form (anchor, positive, negative). Anchor has the same label as positive, whilst negative has a different label. For optimizing the triple loss function, it requires that the feature values of anchor item and positive item extracted from the neural network should as close as possible, whilst the feature values of anchor item and negative item should as far away as possible. However, the image data introduced in the traditional face recognition tasks are always well-structured data with fixed image size and always little difference between human faces [15]. The cross-domain clothing images have always a wide variety of different categories and styles of clothing (such as shirts, jeans, etc.), so the traditional triple loss will not suitable for our cross-domain clothing retrieval task.

Due to above two issues, a novel cross-domain retrieval method of clothing images is proposed in this paper, which is named as ClothingNet and is based upon the feature fusion and quadruplet loss function. The high-level and middle-level features of clothing images can be extracted from our deep neural network and thus can be fused together. The final features of clothing images can be obtained by constraining the feature values to a certain range in term of L_2 norm [16]. Furthermore, in order to overcome the limitations of traditional triple loss function, a novel quadruplet loss function is also introduced here which can be suitable for our clothing retrieval task. The main contributions can be summarized as follows,

- A semantic representation combining high-level and middle-level features of clothing images is presented, which can capture the intrinsic semantic information of garment images comprehensively.
- A new quadruplet loss function is defined, which can be adopted to effectively sense the similarities and differences between different clothing images.
- A cross-domain image retrieval framework based upon the feature fusion and quadruplet loss function is introduced, and the validity and accuracy of our retrieval method can be verified by the public datasets.

II. CLOTHING RETRIEVAL NETWORK WITH FEATURE FUSION AND QUADRUPLET LOSS

The overall framework and the neural network of our cross-domain clothing retrieval is shown in Figure 1. The framework mainly includes two modules, that is, neural network training and image retrieval operation. The network ClothingNet will be stored after network training and employed as a feature extractor for the subsequent cross-domain image retrieval task. Similar to the Siamese network [17], our network comprises two parameter-sharing branches for receiving the input clothing images from consumer domain and shop domain respectively, and each network branch is composed of three units, that is, the basic network unit (in the red box), the feature fusion unit (in the blue box) and the quadruplet loss function unit.

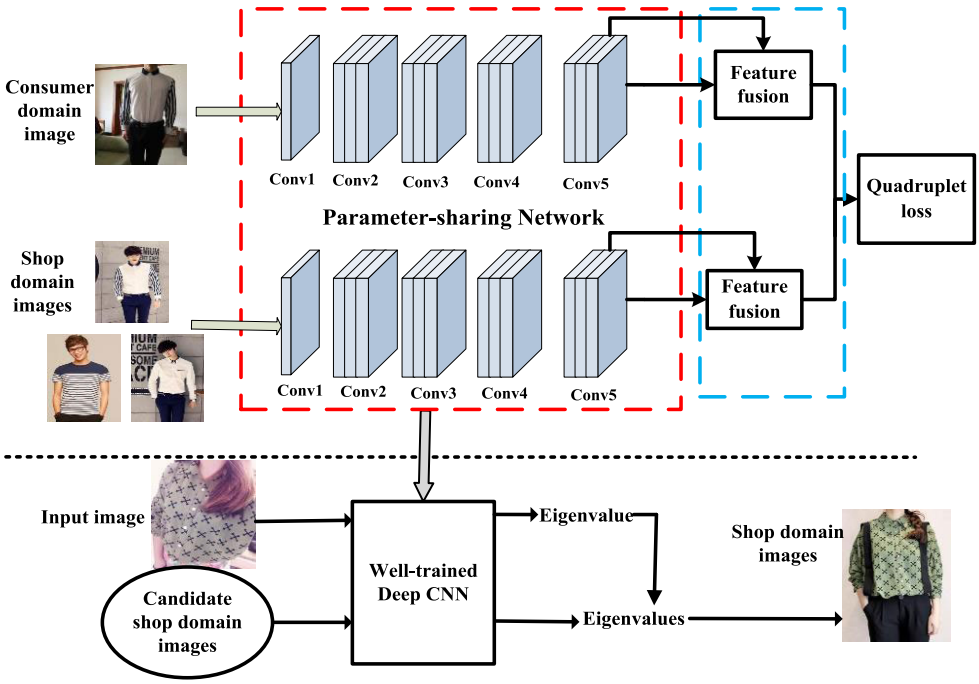


FIGURE 1. Overall framework of our cross-domain clothing image retrieval.

A. BASIC NEURAL NETWORK UNIT

For effectively retrieving clothing images, the key issue is to extract the semantic information of clothing images using multi-layers deep neural networks, such as VGG [18],

GoogleNet [19], ResNet [20] and so on. Since ResNet can effectively overcome the issues of gradient disappearance and data over-fitting of neural network, here the ResNet unit will be adopted as basic network for image feature extraction. ResNet, also known as residual network, has its capability of nonlinear feature expressions due to its introduction of residual blocks. According to the number of network layers, ResNet can be divided into ResNet-18, ResNet-34, ResNet-50 and so on. For our clothing retrieval task, in order to overcome the potential limitations that network errors will expand with the increase of network depth, the original input information will be transmitted directly to the back network layer by setting bypass branches. However, there are some differences between different ResNet networks. For example, for the network ResNet-50, due to the addition of two 1×1 convolution, more convolution layers can thus be superimposed on the same size of receptive fields. The multi-layers convolution means that the neural network can extract abundant feature information of clothing images. Moreover, the 1×1 convolution of the deep neural network can be employed to reduce the latent feature dimension effectively, thus reducing the computational complexity for network training. So, we adopt the ResNet-50 as the basic neural network unit of our presented ClothingNet.

Here, it should be mentioned that the ResNet-50 block contains 50 convolution layers and 2 fully-connected layers. The function of fully-connected layer is to transform the feature graph into a fixed length vector, which can be used as the basis for image classification. Since our retrieval task does not need to classify the underlying images, but only needs to obtain the feature representation of clothing images, our clothing retrieval network ClothingNet will remove the final fully-connected layers which leads to all network parameters introduced by the convolution layers. The network parameters can be listed in Table 1. For our proposed retrieval network, all the parameters between the training network of user domain images and the training network of shop domain images are completely shared. The advantage is that it can not only reduce the number of network parameters and save the computational time for effectively improving the training speed, but also make the neural networks learn more intrinsic sharing characteristics of cross-domain clothing images.

B. FEATURE FUSION UNIT OF CLOTHING IMAGES

For training the proposed network ClothingNet, each fully-connected layer of the network is always a linear classifier. The input of the fully-connected layer is the basis for classification (i.e., the image feature information), and the output of each layer in the network can be adopted as feature representation of the input clothing images. However, the key issue is which layer of the neural network can be selected as the feature representation of clothing images. In reference [4]–[7], image features of the last convolution layer is regarded as feature representation of 2D images. Here, we adopt the output of Conv5-3 (the 3rd convolution layer of Conv5) as the feature representation of clothing images, i.e., image features $f_{conv5-3} \in R^{7\times7\times2048\times N}$ with dimension $7 \times 7 \times 2048 \times N$ as shown in Figure 2. Here, N is the number of training samples

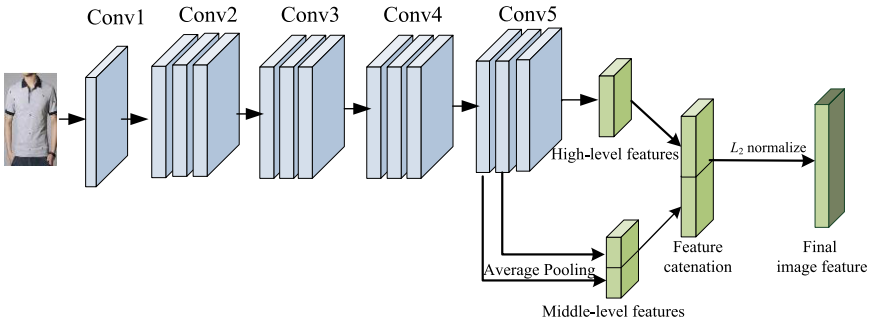


FIGURE 2. Feature fusion of clothing images.

TABLE 1. Network Parameters of Basic Network Unit.

Name	Network type	Size of Kernel	Number of Kernel	Size of Output	Number of Loops
Conv1	Convolution layer	7×7	64	112×112	-
Conv2	Max Pooling layer	3×3	-	56×56	-
	Convolution layer	1×1	64		*3
	Convolution layer	3×3	64		
	Convolution layer	1×1	256		
Conv3	Convolution layer	1×1	128	28×28	*4
	Convolution layer	3×3	128		
	Convolution layer	1×1	512		
Conv4	Convolution layer	1×1	256	14×14	*6
	Convolution layer	3×3	256		
	Convolution layer	1×1	1024		
Conv5	Convolution layer	1×1	512	7×7	*3
	Convolution layer	3×3	512		
	Convolution layer	1×1	2048		

in each batch during network training. However, this scheme for image feature extraction is only suitable for the case which the visual features of retrieval objects are quite different. In the cross-domain retrieval of clothing images, the input images to be searched usually has a large number of common visual characteristics, such as, two different skirts usually have the same length and color, and the similar sleeveless sling skirt may be different only because one have a bow at the waist and the other does not have. Be aimed at these differences of garment images, it is difficult to distinguish the different clothing accurately if only using the high-level semantic information of clothing images.

Here, our proposed ClothingNet network further adopts the middle-level image features which contains the intrinsic local information for the cross-domain clothing retrieval task. In fact, the high-level image features extracted from top layers represent domain-specific features, so we encode the clothing images as a global feature vector and also adopt the global feature representation for clothing retrieval. However, if ignoring the expression of local information of clothing images, it will result in some poor retrieval performance. The middle-level image features extracted from the intermediate layers of neural network could capture the domain-invariant image features.

Vittayakorn *et al.* [11] pointed out that the lower layers of neural network can usually capture the low-level features of images (such as texture, color, etc). And with the increase number of layers of neural network, the captured image feature information becomes more and more abstract. In this paper, we choose the first and second convolution layer Conv5-1 and Conv5-2 of Conv5 to extract the middle-level features of clothing images. For clothing retrieval, the captured clothing images in different domains usually involve different human postures, and even appear at completely different camera angles, such as front, side, back and so on. Therefore, it is difficult to obtain the accurate alignment of image features due to the different camera angles or shooting conditions of captured images. Here, the output eigenvalues of Conv5-1 $f_{conv5-1} \in R^{7\times7\times512\times N}$ and Conv5-2 $f_{conv5-2} \in R^{7\times7\times512\times N}$ in our image retrieval network will be directly employed in the global average pooling operation, which generate the vectors $f_{Pool5-1} \in R^{1\times1\times512\times N}$ and $f_{Pool5-2} \in R^{1\times1\times512\times N}$, respectively. At the same time, these two feature vectors generated by pooling are reduced to $512 \times N$ dimensional vectors which can thus be further catenated and transformed into a $1024 \times N$ dimensional vector. Moreover, the obtained $1024 \times N$ dimensional feature vector can subsequently be reduced to the middle-level image features $f_{middle} \in R^{512\times N}$ by using PCA dimensional reduction algorithm [21]. Meanwhile, the output value $f_{conv5-3} \in R^{7\times7\times2048\times N}$ of Res5-3 can be pooled and the $512\times N$ dimensional vector is also obtained by PCA dimensional reduction algorithm, which can be adopted as the high-level features $f_{high} \in R^{512\times N}$ of clothing images. The obtained middle-level features f_{middle} can thus be connected with the high-level features f_{high} of clothing images. The vector generated by the connection operation can be constrained to the range

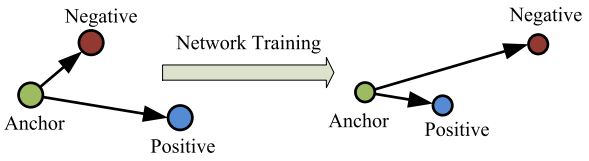


FIGURE 3. Distribution of Anchor, Positive, Negative data.

of [0, 1] by using L_2 norm [16] and the final constrained vector can be considered as the final image feature representation $f_{final} \in R^{1024\times N}$. The detailed pipeline of our feature fusion step is shown in Figure 2.

C. QUADRUPLET LOSS FUNCTION

For the training of neural network, the final clothing image features obtained by feature fusion will be transferred to the loss function module for subsequent calculating and optimizing the network parameters. Before introducing the loss function of our network ClothingNet, the concept of metric learning [22] is introduced here. Metric learning is also known as distance metric learning or similarity learning, which is widely used in the area of computer vision, such as image retrieval, face recognition, etc. In order to achieve the metric learning of training data, the learning samples should be divided into two categories, that is, the same category of samples and different categories of samples [22]. Thus, the key idea of metric learning is to reduce feature distances between the same category of samples as much as possible during the network training, and also to expand the feature distances between different categories of samples as much as possible.

Recently, the loss function that commonly be employed to measure the distance of samples in the literature of metric learning is the triplet loss function, which was first proposed by Schroff *et al.* [14]. The triplet loss function can be constructed as follows. First, a sample Anchor (A for short) is randomly selected from the training set and the corresponding category label Label_A is obtained. Then, a sample Positive (P) is randomly selected from all the data labeled as Label_A. Finally, a sample Negative (N) is randomly selected from all the training data for which label is not Label_A. Thus, the selected samples A, P and N are used to form a triplet group (A, P, N). The relationship between the three categories in this triplet is as follows,

$$\text{Label\_}A = \text{Label\_}P \neq \text{Label\_}N \tag{1}$$

During the network training, the optimization goal of network is to make the sample Anchor and Positive as close as possible in the latent feature space, whilst the sample Anchor and Negative are as far away as possible, which makes the clothing retrieval network better discriminable for different training data as shown in Figure 3.

Inspired by the success of triplet loss applied in the area of face recognition [15], here we consider its application in our cross-domain clothing retrieval task. To be specific, the training data will be organized in the form of triple set Anchor, Positive and Negative, respectively. The Anchor and Positive belong to the same garment, while Negative is the randomly selected clothing image of other clothes category. However, the face image is roughly fixed shape, and its composition is basically unchanged between different human body, which leads to the small difference between different face images [15]. Different from the structured face image data, there are many different categories and styles of clothing and the appearance of different garments could be quite different (such as shirts, jeans, and so on). In this case, after the selection of sample Anchor and Positive, we should choose the sample Negative data. If Negative and Anchor/Positive belong to different garment categories, the distance between Anchor and Negative will be much larger than the distance between Anchor and Positive plus a minimum intervals α. In this case the loss value will not be generated and be considered that the network has reached the ideal state during the network training. So, for all the samples Negative and Anchor, Positive which belong to different garment categories, the neural network may not be effectively trained and could not complete the cross-domain retrieval task. Therefore, in order to effectively perform the cross-domain retrieval of clothing images, we will construct the so-called quadruplet group for the training data and adopt the quadruplet loss function during our neural network training. Quadruplet loss function is another improved version of triplet loss [23]. The quadruplet loss function can be expressed as follows,

$$L_{\text{quad}} = (d_{A,P} - d_{A,N1} + \alpha)_+ + (d_{A,P} - d_{N1,N2} + \beta)_+ \tag{2}$$

The quadruplet loss definition is more conducive to generate effective feature representation of clothing images.

In order to effectively distinguish different garments, in this paper we adopt the hierarchical clothing retrieval strategy. During the image retrieval step, we first determine the category of underlying clothing, and then determine which item belongs to that clothing category. So, each training data of clothing images has double labels, that is, class label and id label. Here, the class label indicates the category of clothing to which the garment belongs, such as T-shirts, shirts, jeans, jackets, jackets, etc. The id label indicates which special garment item that belongs to a particular category. For example, label id_0001 or id_0002 denotes the first or second garment item that belongs to a particular clothing category. Based on the imported class and id labels of clothing images, we constructs a quadruplet (Anchor, Positive, Negative1, Negative2) for training data of clothing images. Unlike the traditional definition of triplet loss, for each Anchor image input from the consumer domain, total 3 clothing images will be selected from the shop domain. They are the Positive image with the same garment id as Anchor, the Negative1 image with the same clothing category but not the same garment id as Anchor, and the Negative2 image with neither the same

FIGURE 4. Example of a cloth quadruplet.

garment id nor the same category as Anchor, as shown in Figure 4.

For the sake of convenience, the quadruplet group is referred to as $(A, P, N1, N2)$ for short below. The images belonging to the same id label is called the same group of clothing images (such as the Anchor and Positive are the same group images), and the images with the same class label is called the same category of clothing images (such as Anchor and Negative1 are the same category images). The traditional triplet loss form can measure the feature distance of training data, which only constrain that the feature distance between same group of clothing images is smaller than the feature distance between different groups of clothing images, i.e., $dist(A, P) < dist(A, N1)$. However, for our cross-domain clothing retrieval task, the hierarchical retrieval strategy should firstly ascertain whether the garment belongs to the same clothing category, and then to determine whether it belongs to the same garment item if it belongs to the same category. Therefore, it is required that the feature distance between similar clothing images should be smaller than that of inter-class images, i.e., $dist(A, N1) < dist(A, N2)$. Therefore, the quadruplet loss function of clothing retrieval needs to satisfy the following constraints. The distance between different clothing categories should be bigger than the distance within the category, and the distance between different garment groups should be bigger than the distance within the group, that is,

$$dist(A, P) < dist(A, N1) \tag{3}$$
$$dist(A, N1) < dist(A, N2) \tag{4}$$

Similar to the distance measurement introduced in the triplet loss function, we also employ Euclidean distance to measure the distance of training data as follows,

$$dist(a, b) = \|f_\theta(a) - f_\theta(b)\|_2 \tag{5}$$

which θ means the parameter set of the neural network, $f_\theta(a)$ and $f_\theta(b)$ represent the fusion features of two clothing images a and b respectively, and $\|\cdot\|_2$ is Euclidean distance.

That is, our proposed quadruplet loss function of ClothingNet needs to satisfy the flowing constraints,

$$\begin{aligned} \|f_\theta(A) - f_\theta(P)\|_2 &< \|f_\theta(A) - f_\theta(N_1)\|_2 \\ \|f_\theta(A) - f_\theta(N_1)\|_2 &< \|f_\theta(A) - f_\theta(N_2)\|_2 \end{aligned} \tag{6}$$

We define here the minimum groups interval between $\|f_\theta(A) - f_\theta(P)\|_2$ and $\|f_\theta(A) - f_\theta(N_1)\|_2$ as m_1, and denote the minimum classes interval between $\|f_\theta(A) - f_\theta(N_1)\|_2$ and $\|f_\theta(A) - f_\theta(N_2)\|_2$ as m_2. Since the differences of image features between clothing categories are usually bigger than that of between garment groups, here $m_1 < m_2$. To sum up, the loss function can be calculated in detail as follows,

$$\begin{aligned} L &= \lambda \cdot L_{id} + \mu \cdot L_{class} \\ &= \lambda \cdot \sum_{i=0}^{N} \left[\|f_\theta(A_i) - f_\theta(P_i)\|_2 + m_1 - \|f_\theta(A_i) - f_\theta(N_{1i})\|_2 \right]_+ \\ &\quad + \mu \cdot \sum_{i=0}^{N} \left[\|f_\theta(A_i) - f_\theta(N_{1i})\|_2 + m_2 \right. \\ &\quad \left. - \|f_\theta(A_i) - f_\theta(N_{2i})\|_2 \right]_+ \end{aligned} \tag{7}$$

III. EXPERIMENTAL RESULTS AND DISCUSSIONS

The proposed network ClothingNet and the pipeline of our cross-domain clothing retrieval have been implemented using the Python language under the development framework of Pytorch deep learning network [24], the running environment is Intel(R) i7-7700, CPU 3.60 GHz, 32G internal memory, and OS Ubuntu 16.04.

A. DATASETS

To verify the effectiveness and accuracy of our clothing retrieval method using feature fusion and quadruplet loss function, we train and test our deep neural network ClothingNet on the following clothing datasets and are also be evaluated with a series of clothes retrieval experiments.

1) DARN dataset [25]: A cross-domain clothing dataset is dedicated to match the street shooting images with online shop clothing images. After the pre-processing step of sorting the image data and deleting the damaged images, a total of 62,812 consumer domain clothing

images and 238,499 shop domain clothing images are obtained, which are the photos of only 13,598 items of different garments. In addition to providing item id labels for each clothing image, the training data also provide several additional labels including categories, colors, lengths, and so on.

2) DeepFashion dataset [4]: A dataset provided by the Multimedia Laboratory of the Chinese University of Hong Kong in 2016. The dataset is the largest clothing image dataset which contains more than 800,000 different categories of clothing images. These clothing images are divided into 4 sub-categories, which can be adopted for the garment attribute prediction, the key-point positioning, the cross-domain clothing retrieval, the clothing alignment and so on. For the issue of cross-domain clothing retrieval task, we use the consumer-to-shop subset. This subset is a set of clothing images aligned between the online store buyer show and the seller show. Each of garment has a separate label id and also have the corresponding one or more images in consumer-to-shop domain that belongs to this label. The dataset contains a total of 239,557 clothing images, including 23 categories and 33,881 clothing item id.

In our experiments, the DARN dataset and the DeepFashion dataset are divided respectively. 70% of the clothing images are selected as the training set, and the rest is divided into 15% test set and 15% verification set.

B. DIFFERENT MTHODS OF CLOTHING RETRIEVAL

In order to validate its effectiveness of our retrieval method, we compare with three traditional cross-domain retrieval methods, such as, DARN [25], WTBI [26], and FashionNet [4]. Moreover, the advantages of our combination of feature fusion and quadruplet loss function are further verified by using the ''feature fusion only'' and the ''quadruplet loss function only'' schemes, respectively.

1) **DARN** benchmark method [25]: For sensing the difference between clothing images from different domains, two different networks with different parameters are used to extract the image features of cross-domain clothing pairs, and the triplet loss function is also adopted.
2) **WTBI** benchmark method [26]: The output of the full connection layer FC6 of the penultimate layer of Alexnet network [27] is directly used as feature representation (4096 dimensional vector) of the clothing images, and the triplet loss function is also adopted.
3) **FashionNet** benchmark method [4]: The attribute information and key-points of the joint clothing are aligned for cross-domain retrieval, and the triplet loss function is also adopted. Since the goal of our method does not involve the positioning operation of key-points, the key-point positioning module in the original neural network is removed.
4) ''**Ours 1**'' method with only feature fusion: Only feature fusion is added to the underlying retrieval network for verifying the effectiveness, and the loss function remains the triplet form.
5) ''**Ours 2**'' method with only quadruplet loss: Only the loss function is modified to the quadruplet loss form for verifying the effectiveness, but the feature representation of the clothing images will not be processed.
6) ''**Ours**'' method combing both feature fusion and quadruplet loss: we add a feature fusion module and modify the loss function to the quadruplet loss form for verifying the effectiveness of our clothing retrieval method.

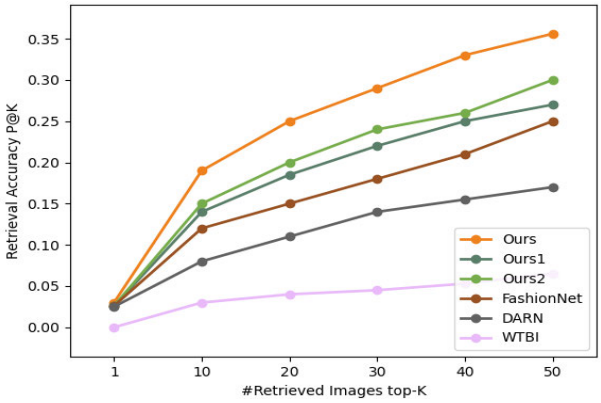


FIGURE 5. Comparisons of retrieval accuracy on DARN dataset.

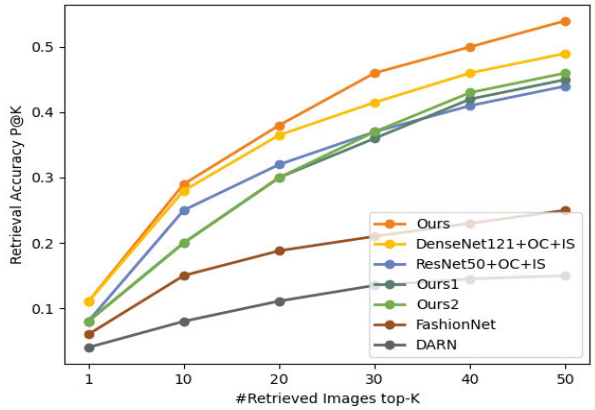


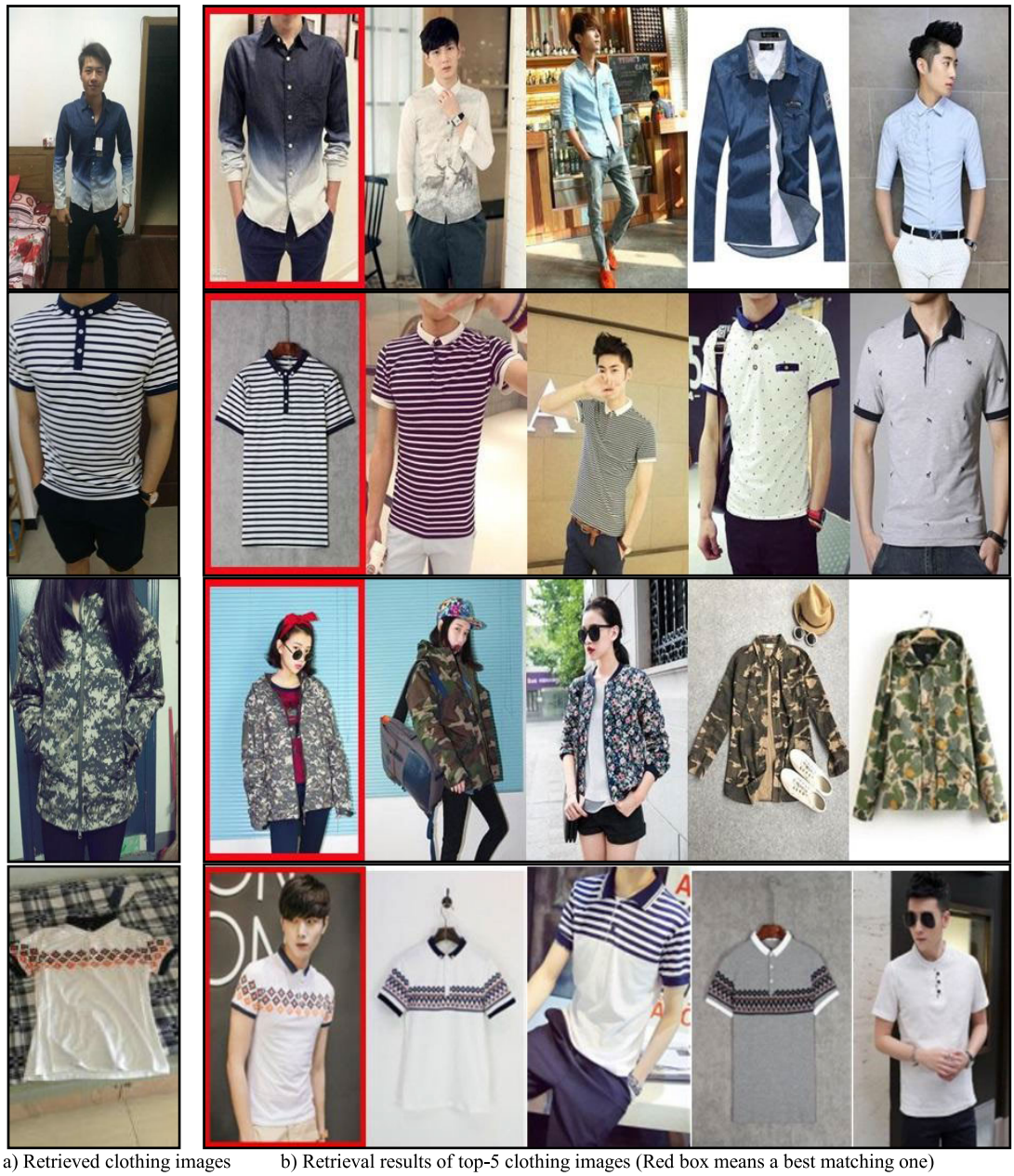
FIGURE 6. Comparisons of retrieval accuracy on DeepFashion dataset.

C. EVALUATION INDEX OF CLOTHING RETRIEVAL

In order to compare the accuracy of cross-domain clothing retrieval, we uses the general top-k accuracy of image retrieval [25] to evaluate the retrieval performance of our method and also benchmark methods, which can be defined as follows,

$$P@K = \frac{\sum_{q \in Q} hit(q, K)}{|Q|} \tag{8}$$

Here, Q means the set of user-captured clothing images to be queried, and q is to a specific image to be queried. If at least one of clothing image contained in the top-k list matches with



a) Retrieved clothing images b) Retrieval results of top-5 clothing images (Red box means a best matching one)

FIGURE 7. Successful retrieval examples of retrieval result in top-5 clothing images.

the image q, the $hit(q, K)$ will be set to 1, otherwise it will be set to 0.

D. EXPRIMENTS OF CLOTHING RETRIEVAL

Before training our presented retrieval network ClothingNet, all the clothing images are standardized to a fixed size of 256*256, and then they will be randomly cropped to 224*224 size clothing images which can be imported into the Restnet-50 network. During the network training, the back propagation algorithm is applied to calculate the gradient of network parameters [28], and the small batch based stochastic gradient descent algorithm is adopted to update the network parameters. During our network training, the mass size is set to 32*32, and the momentum is 0.9. The initial value of the learning rate is set to 0.01, and each 30 epoch is attenuated by 0.1. After the step of network training is completed, the clothing retrieval task can be carried out. The feature vectors of all shop domain images in the test set and the consumer domain images to be searched can thus be extracted by our neural network. The feature distance between the selected consumer domain images and each shop domain image can be calculated respectively, and the shop domain images corresponding to the k smallest distance will be selected as the top-k retrieval results. Similar to DARN [25], WTBI [26], FashionNet [4], we use the k-means algorithm [29] to cluster the feature vectors of all shop domain clothing images for speeding up the retrieval speed.

a) Retrieved clothing images b) Retrieval results of top-5 clothing images

FIGURE 8. Failure retrieval examples of retrieval result in top-5 clothing images.

Figure 5 and Figure 6 show the retrieval accuracy comparisons by using different methods under the DARN dataset [25] and DeepFashion dataset [4]. According to Figure 5 and Figure 6, it can be seen that the retrieval accuracy of different methods on DeepFashion dataset is better than that on DARN dataset. This is because of the difference between these two datasets, for example, the size of DeepFashion dataset is much larger than that of DARN dataset, and the variety and quantity of clothes are also bigger than that of DARN dataset. So, the accuracy of clothing retrieval task on the DARN dataset is always lower than that of DeepFashion dataset.

As also can be seen from Figure 5 and Figure 6, the retrieval performance of WTBI method [26] on both datasets is the worst among the six retrieval schemes. It is because WTBI method directly adopts the output of final fully-connected layer of network as feature representation of clothing images, which is a 4096 dimensional feature vector. Large of image feature dimension makes the large scale of overall network computation, which will lead to data mis-fitting and also a low retrieval accuracy. The retrieval accuracy of FashionNet method [4] is better than that of DARN [25] and WTBI [26], which is consistent with the experimental results in reference [26]. Compared with three traditional benchmark methods, it can be seen that the accuracy of ''Ours'' clothing retrieval method is obviously improved after using feature fusion or quadruplet loss, and the accuracy improvement caused by the feature fusion is more obviously than that caused by the quadruplet loss. Moreover, ''Ours'' method which combing both feature fusion and quadruplet loss will perform the best one among six retrieval methods. On the DARN and DeepFashion datasets, we take the top-50 retrieval results for our comparisons. The retrieval accuracy of ''Ours'' method are 35.67% and 53.52%, respectively, which are almost 10% and 11% higher than those of traditional benchmark methods, respectively. Meanwhile, the retrieval accuracy of ''Ours'' method on the DeepFashion dataset is also higher than that of DenseNet121+OC+IS and ResNet50+OC+IS [10]. Experimental results can verify the effectiveness and accuracy of our proposed clothing retrieval method.

Figure 7 shows several successful retrieval examples of our image retrieval method by using ClothingNet. Figure 7a is the user input clothing images to be retrieved, and five of the clothing images listed in Figure 7b are the most matched shop domain images retrieved by our method, wherein the red frame marks the most matching one. It can be seen from Figure 7 that the retrieval accuracy of clothing images is relatively high for the query images with remarkable characteristics, such as lines with specific laws, lattices and various printing patterns, if providing the front-side photos and sufficient shooting light for capturing clothing images. Due to its ability of rich feature extraction, the deep neural network can learn rich feature embedding and exceed the manual extracted image features, such as CGTW and CGOT [30].

Figure 8 gives two failure examples of our clothing retrieval method, that is, it can not search the exact matching clothes among the returned top-5 retrieval results due to its poor quality of photos taken by the user. Here, the input clothing image of the first row is a check shirt image taken from the side which leads to some deformation artifacts. Our presented neural network can only capture the local pattern and texture information of input image, but it is difficult to extract the global information of clothing image, which leads

to some retrieval results of similar shirts with different texture print patterns. The second input image is a clothing image of jeans and leads to some retrieval results of roughly similar but different jeans. It is due to its far away the shooting distance and the camera can not effectively focus on the target regions.

In summary, if providing the clothing images that consumers take under regular shooting conditions and practical scenarios, our clothing retrieval method with feature fusion and quadruplet loss is always effective, and the retrieval accuracy is better than that of the existing traditional methods.

IV. CONCLUSION

In order to overcome the issues of low accuracy existing in traditional cross-domain clothing retrieval methods, based upon the image feature fusion and quadruplet loss function, a novel retrieval method of clothing images is presented in this paper by using network ClothingNet. Different from other existing retrieval methods, our feature representation of clothing images is not limited to the final high-level features of convolution network, but the fusion of middle-level features and high-level features. Moreover, the quadruplet loss function will combine the category and style attributes of clothing for training the underlying neural network. Comparative experiments have been carried out on the public datasets for cross-domain clothing retrieval, such as DARN and DeepFashion. Experimental results illustrate the effectiveness of our clothing retrieval method, and the retrieval accuracy is always superior to the existing schemes.

However, if object occlusions or image deformation artifacts of clothing images captured by the users is serious, the retrieval results maybe failure and not be satisfactory. The future work will consider how to achieve high precision retrieval effects in the case of large area occlusions and deformations while obtaining the clothing images.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their helpful and valuable comments. They also appreciate the availability of public DARN dataset and DeepFashion dataset for our experiments. Prof. Yongwei Miao is a Qianjiang Special Expert of Hangzhou city at Hangzhou Normal University.

REFERENCES

[1] S. Liu, Z. Song, G. Liu, C. Xu, H. Lu, and S. Yan, "Street-to-shop: Cross-scenario clothing retrieval via parts alignment and auxiliary set," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 3330–3337.

[2] Y. Kalantidis, L. Kennedy, and L.-J. Li, "Getting the look: Clothing recognition and segmentation for automatic product suggestions in everyday photos," in *Proc. 3rd ACM Conf. Int. Conf. Multimedia Retr. (ICMR)*, Apr. 2013, pp. 105–112.

[3] W. Liu, Z. Wang, X. Liu, N. Zeng, Y. Liu, and F. E. Alsaadi, "A survey of deep neural network architectures and their applications," *Neurocomputing*, vol. 234, pp. 11–26, Apr. 2017.

[4] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang, "DeepFashion: Powering robust clothes recognition and retrieval with rich annotations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1096–1104.

[5] B. Gajic and R. Baldrich, "Cross-domain fashion image retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 1869–1871.

[6] Y. Xiong, N. Liu, Z. Xu, and Y. Zhang, "A parameter partial-sharing CNN architecture for cross-domain clothing retrieval," in *Proc. Vis. Commun. Image Process. (VCIP)*, Nov. 2016, pp. 1–4.

[7] X. Ji, W. Wang, M. Zhang, and Y. Yang, "Cross-domain image retrieval with attention modeling," in *Proc. ACM Multimedia Conf. (MM)*, 2017, pp. 1654–1662.

[8] Y. Luo, Z. Wang, Z. Huang, Y. Yang, and H. Lu, "Snap and find: Deep discrete cross-domain garment image retrieval," 2019, *arXiv:1904.02887*. [Online]. Available: http://arxiv.org/abs/1904.02887

[9] M. Kucer and N. Murray, "A detect-then-retrieve model for multi-domain fashion item retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 344–353.

[10] S. Park, M. Shin, S. Ham, S. Choe, and Y. Kang, "Study on fashion image retrieval methods for efficient fashion visual search," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 316–319.

[11] S. Vittayakorn, T. Umeda, K. Murasaki, K. Sudo, T. Okatani, and K. Yamaguchi, "Automatic attribute discovery with neural activations," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Oct. 2016, pp. 252–268.

[12] D. C. Ciresan, U. Meier, J. Masci, L. M. Gambardella, and J. Schmidhuber, "Flexible, high performance convolutional neural networks for image classification," in *Proc. 22th Int. Joint Conf. Artif. Intell. (IJCAI)*, Jul. 2011, pp. 1237–1242.

[13] M. Uličný, J. Lundström, and S. Byttner, "Robustness of deep convolutional neural networks for image recognition," in *Intelligent Computing Systems*. Cham, Switzerland: Springer, 2016.

[14] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 815–823.

[15] A. S. Tolba, A. H. El-Baz, and A. A. El-Harby, "Face recognition: A literature review," *Int. J. Signal Process.*, vol. 2, no. 2, pp. 88–103, 2006.

[16] R. Ranjan, C. D. Castillo, and R. Chellappa, "L_2-constrained softmax loss for discriminative face verification," 2017, *arXiv:1703.09507*. [Online]. Available: http://arxiv.org/abs/1703.09507

[17] I. Melekhov, J. Kannala, and E. Rahtu, "Siamese network features for image matching," in *Proc. 23rd Int. Conf. Pattern Recognit. (ICPR)*, Dec. 2016, pp. 378–383.

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

[19] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[21] S. Roweis, "EM algorithms for PCA and SPCA," in *Proc. Int. Conf. Neural Inf. Process. Syst. (NIPS)*, Jul. 1997, pp. 626–632.

[22] E. P. Xing, M. I. Jordan, S. J. Russell, and A. Y. Ng, "Distance metric learning with application to clustering with side-information," in *Proc. Int. Conf. Neural Inf. Process. Syst. (NIPS)*, Jun. 2003, pp. 521–528.

[23] W. Chen, X. Chen, J. Zhang, and K. Huang, "Beyond triplet loss: A deep quadruplet network for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1320–1329.

[24] N. Ketkar, *Deep Learning With Python: A Hands-On Introduction*. Berkeley, CA, USA: Apress, 2017, pp. 195–208.

[25] J. Huang, R. Feris, Q. Chen, and S. Yan, "Cross-domain image retrieval with a dual attribute-aware ranking network," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1062–1070.

[26] 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 *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3343–3351.

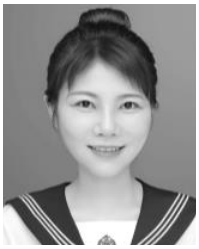
[27] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Int. Conf. Neural Inf. Process. Syst. (NIPS)*, Dec. 2012, pp. 1097–1105.

[28] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, May 2015.

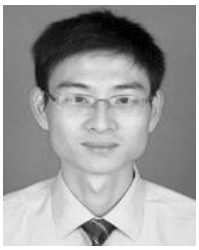
[29] J. A. Hartigan and M. A. Wong, "Algorithm AS 136: A k-means clustering algorithm," *Appl. Statist.*, vol. 28, no. 1, pp. 100–108, 1979.

[30] Z. Shao, W. Zhou, L. Zhang, and J. Hou, "Improved color texture descriptors for remote sensing image retrieval," *J. Appl. Remote Sens.*, vol. 8, no. 1, pp. 083584:1–083584:13, Jul. 2014.

YONGWEI MIAO (Member, IEEE) received the master's degree in mathematics from the Institute of Mathematics, Chinese Academy of Sciences, in July 1996, and the Ph.D. degree in computer graphics from the State Key Laboratory of CAD and CG, Zhejiang University, in March 2007. From February 2008 to February 2009, he was a Visiting Scholar with the University of Zurich, Switzerland. From November 2011 to May 2012, he was a Visiting Scholar with the University of Maryland, USA. From July 2015 to August 2015, he was a Visiting Professor with The University of Tokyo, Japan. He is currently a Professor with the College of Information Science and Technology, Zhejiang Sci-Tech University, and also a Visiting Professor with the College of Information Science and Engineering, Hangzhou Normal University. He has authored or coauthored more than 125 technical articles published in scientific journals or presented at conferences. His research interests include computer graphics, digital geometry processing, 3D computer vision, visual media computing, and deep learning.



GAOYI LI received the master's degree from the College of Computer Science and Technology, Zhejiang University of Technology, in June 2019. Her research interests include computer graphics, computer vision, and deep learning.



CHEN BAO is currently pursuing the Ph.D. degree with the College of Computer Science and Technology, Zhejiang University of Technology. His research interests include computer graphics, computer vision, interactive modeling and design of 3D garments, and deep learning.

JIAJING ZHANG received the Ph.D. degree in computer graphics from the State Key Laboratory of CAD and CG, Zhejiang University, in January 2018. She is currently a Faculty Member with the College of Information Science and Technology, Zhejiang Sci-Tech University. Her research interests include computer graphics, computational aesthetics, and visual media computing.

JINRONG WANG received the master's degree in cryptography theory and practices from Hangzhou Dianzi University, in June 2004, and the Ph.D. degree in computer graphics from the State Key Laboratory of CAD and CG, Zhejiang University, in 2012. He is currently a Faculty Member with the College of Information Science and Engineering, Hangzhou Normal University. His research interests include computer graphics, digital geometry processing, 3D shape modeling, digital watermarking, and 3D computer vision.