



## Article

# Person Re-Identification by Low-Dimensional Features and Metric Learning

Xingyuan Chen <sup>1,\*</sup>, Huahu Xu <sup>1,2</sup>, Yang Li <sup>1</sup> and Minjie Bian <sup>1,2</sup>

<sup>1</sup> School of Computer Science and Engineering, Shanghai University, Shanghai 200444, China; huahuxu@staff.shu.edu.cn (H.X.); liyang162@shu.edu.cn (Y.L.); bianmj0302@aliyun.com (M.B.)

<sup>2</sup> Shanghai Shangda Hairun Information System Co., Ltd., Shanggai 200072, China

\* Correspondence: cccxy@shu.edu.cn

**Abstract:** Person re-identification (Re-ID) has attracted attention due to its wide range of applications. Most recent studies have focused on the extraction of deep features, while ignoring color features that can remain stable, even for illumination variations and the variation in person pose. There are also few studies that combine the powerful learning capabilities of deep learning with color features. Therefore, we hope to use the advantages of both to design a model with low computational resource consumption and excellent performance to solve the task of person re-identification. In this paper, we designed a color feature containing relative spatial information, namely the color feature with spatial information. Then, bidirectional long short-term memory (BLSTM) networks with an attention mechanism are used to obtain the contextual relationship contained in the hand-crafted color features. Finally, experiments demonstrate that the proposed model can improve the recognition performance compared with traditional methods. At the same time, hand-crafted features based on human prior knowledge not only reduce computational consumption compared with deep learning methods but also make the model more interpretable.



**Citation:** Chen, X.; Xu, H.; Li, Y.; Bian, M. Person Re-Identification by Low-Dimensional Features and Metric Learning. *Future Internet* **2021**, *13*, 289. <https://doi.org/10.3390/fi13110289>

Academic Editor: Maria Gabriella Xibilia

Received: 4 November 2021  
Accepted: 16 November 2021  
Published: 18 November 2021

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

**Keywords:** person re-identification; hand-crafted features; color feature; BLSTM

## 1. Introduction

Person re-identification (Re-ID) is a popular direction in computer vision research, with a wide range of application scenarios, such as real-time monitoring, trajectory tracking, security and other applications. The core of the task is to judge whether a pedestrian's different images captured in a non-overlapping field of view belong to the same person. However, there are different viewpoints, illumination changes and other complex factors, which make the Re-ID problem difficult to solve.

Most of the traditional methods, which focused on metric learning [1–3] and feature representation [4,5], are commonly used to extract features from the whole image or manually designed horizontal windows. The renaissance of deep learning has dominated this field. Most studies [6–8] use convolutional neural networks (CNNs) to extract global or local features combined with metric learning losses to solve problems. However, the features extracted by the CNN can easily ignore both the spatial structure of the person and attribute features, and sometimes it is precisely these features that play a decisive role in identifying pedestrians and can give the model better interpretability. Therefore, there are some studies that integrate human structural information [9] into tasks, and some studies that introduce attention mechanisms (AMs) [10,11], which can focus on important points in substantial information, selecting key information and ignoring other unimportant information, to obtain pedestrian attributes. These methods have achieved good results. However, due to low-image resolution, illumination changes, unconstrained poses and different viewpoints, the local features or global features displayed by pedestrians exhibit large differences, which may allow neural networks to learn useless features, so there is still a large gap between research-oriented scenarios and practical applications [12].

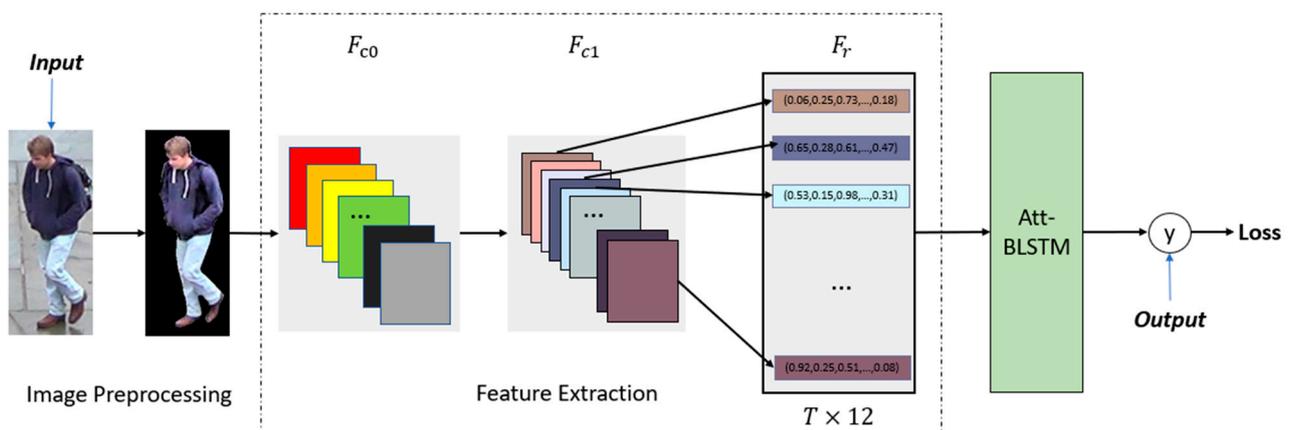
Color feature is one of the most commonly used features in traditional methods, which mostly extract color features from the “hue, saturation, value (HSV)” color space. However, due to the illumination variation and lack of spatial information, the traditional methods are not robust. Although most recent studies ignore color features and focus on deep features, color features still have an important research significance for person re-identification tasks regarding low-resolution images; most person pictures have a resolution of less than  $300 \times 300$ , as shown in Table 1, and contain much less information about a person, which may make the neural network learn many useless features that not only affect the computational efficiency but also cause a waste of resources.

Therefore, in this paper we propose a color feature with spatial information, which include not only person HSV color space features, but also space information. Based on the color feature, we implement bidirectional long short-term memory (BLSTM) networks with an attention mechanism to obtain the contextual relationship contained in the spatial color feature. The experimental results demonstrate that the proposed model can improve the recognition performance compared with traditional methods. At the same time, hand-crafted features based on human prior knowledge not only reduce computational consumption compared with deep learning methods but also make the model more interpretable. An overview of our proposed model is shown in Figure 1. The main contributions and novelty of this paper are as follows:

- Based on the prior knowledge of humans, we designed a color feature with spatial information to solve the problem of person re-identification. Compared with the common method of extracting deep features using convolutional neural networks, the biggest advantage of the handcraft features we designed is that it consumes less computing resources during extraction, has better interpretability, and is less affected by image resolution.
- Att-BLSTM is used to obtain the contextual semantic relationship in the color features, and due to the attention mechanism, the model can automatically focus on the features that are decisive for the task. The performance of the model and its generalization ability can be greatly improved.
- The combination of hand-crafted features and the deep learning in person re-identification task not only greatly reduces the number of parameters and the resource consumption of training models, but also gives the model better interpretability, and the performance of the model can still reach an advanced level.

**Table 1.** Image resolution of common data sets for person re-identification.

Datasets	Time	Images	Resolution
VIPeR	2007	1264	$128 \times 48$
CUHK01	2012	3884	$160 \times 60$
Market-1501	2015	32,668	$128 \times 64$
MARS	2016	1,191,003	$256 \times 128$
DukeMT-MCrelD	2017	36,441	vary



**Figure 1.** The framework of the model. The output  $y$  will be used to calculate the hard triplet loss, which will be mentioned later.

## 2. Related Works

### 2.1. Person Re-Identification

Person Re-ID judges whether pedestrians in different images captured in a non-overlapping field of view belong to the same person. Therefore, the low resolution of pedestrian images, the diversity of portrait shooting angles and pedestrian postures are the main problems faced by this task [13]. Early research has mainly been focused on hand-craft feature extraction [3,14,15] and metric learning methods [16], which still have much room for improvement. Li et al. [4] combined spatial information with low-level features and proposed a feature representation method called local maximal occurrence (LOMO). Layne et al. [5] also used pedestrian attribute characteristics to re-identify pedestrians, and they designed and manually labeled 15 pedestrian attributes based on low-level characteristics for clothing style, hairstyle, personal belongings, and gender. Recently, deep learning has become unique in the field of pedestrian dynamics, and many research advances have been made [9,11,17,18]. In [17], the authors proposed using a generative adversarial network (GAN) to generate unlabeled samples to improve the baseline. Wang et al. [9] regarded the different local features of pedestrians as different nodes of the graph to learn relational information and then used graph matching algorithms to enhance the robustness of the model. Qian et al. [16] proposed a multiscale deep representation learning model to capture distinguishing features at different scales.

Although the ability of convolutional neural networks to extract visual features is widely recognized, the biggest difference between the person re-identification task and other computer vision tasks is the resolution of images, as shown in Table 1. A low-resolution image contains much less information about a person, which may make the neural network learn many useless features, and at the same time makes the computational resources required to train the model larger.

### 2.2. Color Feature

The color feature is one of the main features in person re-identification. Color has enormous value in recognition because it is a local surface property that is view invariant and largely independent of resolution [19]. Many studies have already used color features for person re-identification. Hu et al. [20] proposed the weighted color topology (WCT) feature is proposed to exploit the spatial distribution information. In [21], the authors propose a feature fusion method for person re-identification, which includes the HSV and color histogram features and the texture feature extracted by the HOG descriptor. In [22], the authors believe that the same feature extraction and description of all parts without differentiating their characteristics will result in poor re-identification performances, and propose an algorithm to fully exploit region-based feature salience.

The biggest problem with handcraft color features is that if we do not use other tools, we can only use the information it represents, but we cannot obtain the high-level semantic features such as the contextual relationship and semantic information contained in the color feature. Therefore, we combine the handcraft color feature with the deep learning model, which can discover the high-level semantic features implied by the color feature, to solve the task of person re-identification.

### 2.3. Recurrent Neural Network

A recurrent neural network (RNN) is a type of deep neural network that has recurrent connections, which enables the network to capture the context information in the sequence and retain the internal states. To overcome the vanishing gradient problem of the RNN, long short-term memory (LSTM) units were introduced by Hochreiter and Schmidhuber. The main idea is to introduce an adaptive gating mechanism, which decides the degree to which the LSTM units keep the previous state and memorize the extracted features of the current data input. As a result, they may be better at finding and exploiting long-range dependencies in the data. To date, many LSTM variants have been proposed, such as BLSTM and bidirectional long short-term memory networks with an attention mechanism (Att-BLSTM) [23].

In the field of natural language, LSTM plays an increasingly important role, and an increasing number of researchers have been applying LSTM to computer vision applications. Li et al. [24] used LSTM to mine the semantic correlation and spatial information of pedestrian attributes to improve the performance of pedestrian attribute recognition. In [25], the authors used recursive LSTM to generate sequences and train the model end-to-end to solve people detection in crowded scenes. In [26], the authors proposed a novel Siamese LSTM architecture that can process image regions sequentially and enhance the discriminative capability of the local feature representation by leveraging contextual information.

## 3. The Proposed Method

### 3.1. Image Preprocessing

Color is quite sensitive to illumination changes; therefore, the lighting conditions of different cameras or camera settings will inevitably lead to differences in the imaging colors of the same pedestrian. To solve this problem, as in [4], we use the Retinex algorithm [27] to preprocess images. In addition, we automatically compute the gain/offset parameters so that the resulting intensities linearly stretch in (0, 255). The algorithm enhances the details of the shadow area of the original images so that the color of each picture is as vivid as that observed by a human, as shown in Figure 2b.



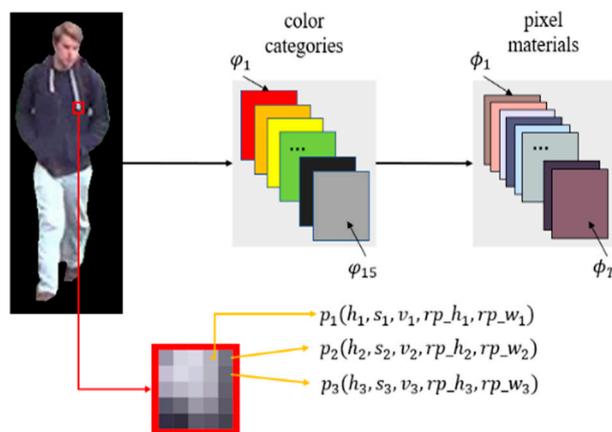
**Figure 2.** Examples of image preprocessing process. (a) Example input images from the DukeMTMC-reID database. (b) Processed images by the Retinex algorithm. (c) Images after preprocessing.

Additionally, eliminating the negative effects of background noise is equally important. With the continuous innovation of image segmentation algorithms based on deep learning, including mask region-based convolutional neural network (RCNN) [28], DeepLab v3 [29], dense pose [30], etc., an increasing number of works [6,12] have introduced semantic segmentation into Re-ID. In our work, the Deeplab v3 model is used to obtain the range of interest (ROI). Therefore, the ROI can be obtained after the original image is preprocessed, as shown in Figure 2c.

### 3.2. Hand-Crafted Feature Extraction

The image resolution of person Re-ID images is quite low (mostly less than  $300 \times 300$ ); these images contain much less information than those of normal resolution, and many appearance details are lost [31]. Therefore, color becomes the most important feature for describing person images. At the same time, the classification of colors is a manifestation of human knowledge, and compared with other characteristics, colors are not easily affected by changes in posture and viewing angle, so they have relatively better robustness in person Re-ID. In this section, we will introduce in detail the color feature with spatial information proposed in our model.

Based on the bag of feature algorithm (BoF) [32–34], where the main idea is to quantify the disordered set of local descriptors, we use another widely used color space, the so-called “hue, saturation, value (HSV)” color space, instead of the “red, green, blue (RGB)” color space to sharpen color images. We divide the HSV color space into 15 color categories, including three achromatic colors (black, white, gray) and twelve colors (blue, green, yellow, etc.). Then, as shown in Figure 3, the basic features of all pixels in the ROI are extracted in turn and classified according to 15 colors.

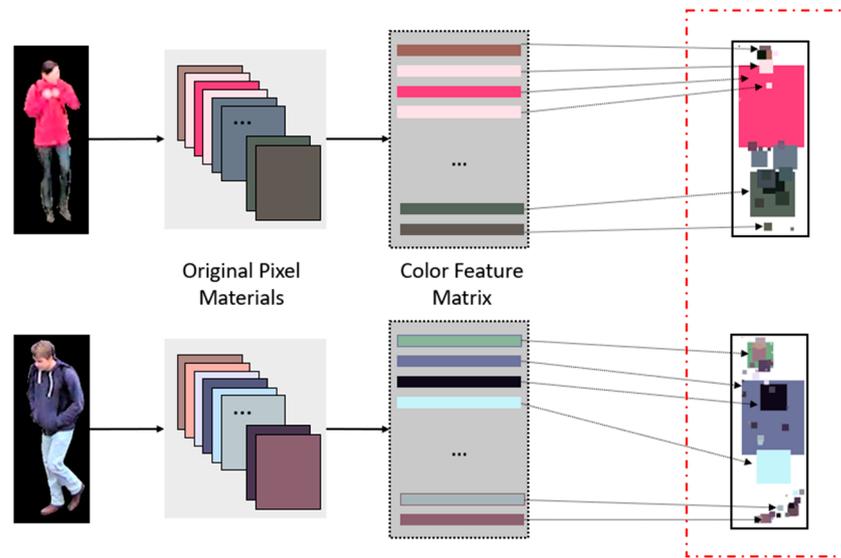


**Figure 3.** The pixel material extraction process description.  $h_i, s_i, v_i, rp_{h_i}, rp_{w_i}, i = 1 \dots N$  indicate hue, saturation, value of the pixel, height distance from the origin, and width distance from the origin of the pixel, respectively.  $N$  is the total number of ROI pixels.

For all pixels in each color category  $\varphi = \{p_1, p_2, \dots, p_s\}$ , if the longitudinal relative difference between two adjacent pixels exceeds the threshold  $T_h$ , then the set is divided into  $\phi_1 = \{p_1, p_2, \dots, p_i\}, \phi_2 = \{p_{i+1}, p_{i+2}, \dots, p_s\}$ . After several longitudinal splits, we obtain  $\varphi_h = \{\phi_1, \phi_2, \dots, \phi_{h+1}\}$ , where  $h$  is the number of times that adjacent pixels exceed the threshold. Later, each set in the set  $\varphi_h$  is divided horizontally. Finally,  $\varphi_w = \{\phi_1, \phi_2, \dots, \phi_{t+1}\}$ , where  $t$  is the sum of the number of vertical and horizontal divisions. Fifteen color categories are separated in turn and finally  $T$  original pixel materials are obtained, namely,  $\varphi_T = \{\phi_1, \phi_2, \dots, \phi_T\}$ .

To allow the  $T$  original pixel materials obtained in the previous section to fully express the information they contain, we convert each original pixel material in  $\varphi_T$  into a feature vector, as shown in Figure 4. Each vector is 14-dimensional, which is the variance and mean of the  $H, S,$  and  $V$  of all pixels in the original pixel material, and the ratio of the number of

original pixel materials to all pixels in the ROI. In addition, relative spatial information is incorporated, as shown in Figure 4. Each pixel includes not only the HSV color feature but also the spatial position information of the pixel relative to the origin, which is the lower left corner of the image. Then, we can calculate the relative spatial information contained in the original pixel materials.



**Figure 4.** Visual description examples of the hand-crafted color feature matrix. The images within the red dashed line are the visual descriptions of the matrix generated by the original images.

Taking the longitudinal relative spatial information as an example, the average relative height ( $AR_h$ ) and relative height standard deviation ( $SDR_h$ ) are defined as (2) and (3), respectively.

$$rp\_h_i = \frac{Pixel_i - RoI_{min}}{RoI_{max} - RoI_{min}} \tag{1}$$

$$AR_h = \frac{\sum_i^n rp\_h_i}{n} \tag{2}$$

$$SDR_h = \sqrt{\frac{\sum_i^n (rp\_h_i - AR_h)^2}{n - 1}} \tag{3}$$

### 3.3. Att-BLSTM

Based on the color features with spatial information, which obtained in the previous section. We chose Att-BLSTM networks to obtain the contextual semantic relationship and the most important semantic information in the color features. Att-BLSTM uses BLSTM networks to obtain contextual semantic information in sentences, and the AM is combined with BLSTM so that it can automatically focus on the features that are decisive for the task. As the color features can be regarded as the feature after word embedding, we made some modifications to Att-BLSTM, removing the word embedding layer, as shown in Figure 5.

Generally, the LSTM-based RNNs consist of the following components: one input gate  $i_t$  with corresponding weight matrix  $W_{xi}, W_{hi}, W_{ci}, b_i$ ; one forget gate  $f_t$  with corresponding weight matrix  $W_{xf}, W_{hf}, W_{cf}, b_f$ ; one output gate  $o_t$  with corresponding weight matrix  $W_{xo}, W_{ho}, W_{co}, b_o$ . All of those gates are set to generate some degrees using the current input  $x_i$ , the state  $h_{i-1}$  that the previous step generated, and the current state of this cell  $c_{i-1}$ . For the decisions whether to take the inputs, forget the memory stored before, and output the state generated later [23]. The components are demonstrated in (4)–(9) as follows.

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i) \tag{4}$$

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f) \tag{5}$$

$$g_t = \tanh(W_{xc}x_t + W_{hc}h_{t-1} + W_{cc}c_{t-1} + b_c) \tag{6}$$

$$c_t = i_t g_t + f_t c_{t-1} \tag{7}$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_{t-1} + b_o) \tag{8}$$

$$h_t = o_t \tanh(c_t) \tag{9}$$

where  $\sigma$  is the activation function;  $i, f, o$  are the input gate, forget gate, output gate, and current cell state, respectively; and  $c_t$  can be calculated by the weighted sum using both the previous cell state and current information generated by the cell.

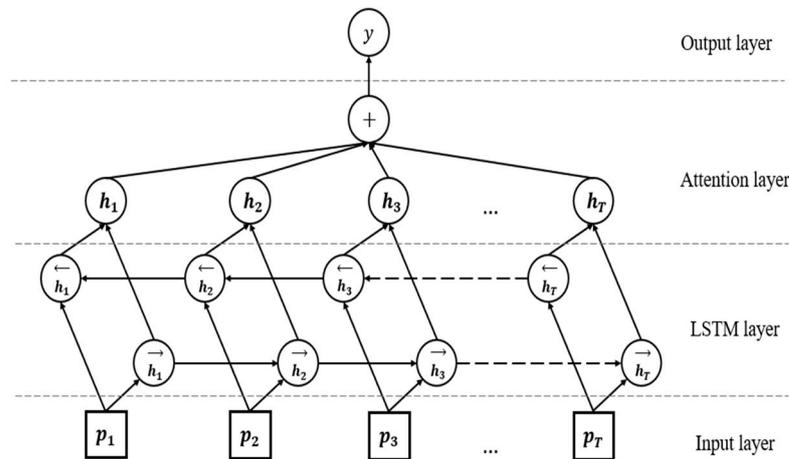


Figure 5. Bidirectional LSTM model with attention.

Additionally, as shown in Figure 5, Att-BLSTM contains two subnetworks for the left and right sequence contexts, which are forward and backward passes, respectively. The output of the  $i$ -th word is shown in the following equation:

$$h_i = \left[ \begin{array}{c} \vec{h}_i \oplus \overleftarrow{h}_i \end{array} \right] \tag{10}$$

Then, we introduce the AM to obtain abstract representation, as the following equations demonstrate:

$$M = \tanh(H) \tag{11}$$

$$\alpha = \text{softmax}(Mw^T) \tag{12}$$

$$\gamma = \alpha^T H \tag{13}$$

where  $H = [h_1, h_2, \dots, h_T]$  is the output of BLSTM and  $T$  is the RR length.  $d^w$  is the dimension of the word vectors,  $w$  is a trained parameter vector, and  $w^T$  is a transpose.  $\gamma$  can be formed by a weighted sum of these output vectors. The dimensions of  $w, \alpha$ , and  $\gamma$  are  $d^w, T$ , and  $d^w$ , respectively. Finally, we can obtain the abstract representation of pedestrian  $H^*$  as (14):

$$H^* = \tanh(\gamma) \tag{14}$$

### 3.4. Loss Function

As in [8,35], we apply the triplet hard loss [36] as the loss of metric learning to improve the generalization ability and training speed of the model. The process is as follows. First, randomly select  $P$  pedestrians in the training dataset, randomly select  $K$  pictures for each pedestrian, and obtain a total of  $PK = X$  pictures as a batch. Input these  $X$  pictures into

the model to obtain  $X$  groups of pedestrian abstract representations, which can be used to calculate the loss of pedestrians using the following equation:

$$L_{id}(\theta; X) = \sum_{i,j=1}^P \sum_{a=1}^K \left[ \max_{p=1, \dots, K} d(H_a^{*(i)}, H_p^{*(i)}) - \min_{n=1, \dots, K} d(H_a^{*(i)}, H_n^{*(j)}) + s \right] \quad (15)$$

where  $H_a^{*(i)}$  represents the abstract representation feature of the  $a$ -th image of the  $i$ -th pedestrian and uses this as the anchor.  $H_p^{*(i)}$  and  $H_n^{*(j)}$  represent the abstract representation features of the positive pedestrian and the negative pedestrian corresponding to the anchor, respectively, and  $s$  is an artificially set hyperparameter. The abstract representation feature similarity  $d(H_i^*, H_j^*)$  of any two pedestrians  $H_i^*$  and  $H_j^*$  can be calculated by the following equation:

$$d(H_i^*, H_j^*) = \frac{H_i^* \cdot H_j^*}{\|H_i^*\| \times \|H_j^*\|} \quad (16)$$

## 4. Experiments and Discussion

To evaluate the proposed model, we selected two large-scale person Re-ID benchmarks, DukeMTMC-reID [17] and Market1501 [37].

### 4.1. Datasets and Evaluation Protocol

#### 4.1.1. DukeMTMC-reID

DukeMTMC-reID is a subset of the DukeMTMC dataset for image-based Re-ID proposed by Duke University. It consists of 36,411 images of 1812 identities collected by eight different cameras. A total of 16,522 images of 702 persons are divided into a training set. A total of 19,889 images of the remaining identities are divided into a testing set, with 2228 in the query set and 17,661 in the gallery set.

#### 4.1.2. Market-1501

The Market-1501 dataset was constructed and made public in 2015, and the dataset was collected by six cameras. There are 1501 different person identities in the dataset. They use the DPM detector to generate the person detection frame, and get 32,668 pedestrian pictures. These pictures are divided into non-overlapping training set and test set. The training set contains 12,936 pictures corresponding to 751 persons, and the test set contains 19,732 pictures corresponding to 750 persons. In addition, the person detection frames of 3368 query pictures are drawn manually to ensure the clarity of the target to be queried during the test.

#### 4.1.3. Evaluation Protocol

The common criteria for evaluating the performance of person re-identification algorithms include cumulative matching characteristics (CMC) and mean average precision (mAP) [13]. CMC- $N$  (i.e., Rank- $N$  matching accuracy) means the probability that the highest confidence  $n$  images in the search results have the correct results. Another metric, i.e., mAP is originally widely used in image retrieval. For Re-ID evaluation, the essence of mAP is actually an average value of the maximum recall rate of each person in multi-person detection. The Rank-1 and mAP results are reported. All the experiments are performed in a single query setting

### 4.2. Implementation Details

We implement our framework with Pytorch. The specific experimental environment is as follows: Python 3.6.7, CUDA 10.0, and CUDNN 7.5.

As in [8], the margins of triplet hard loss for cosine distances are set to 0.5, and the minibatch size is set to 160, in which 4 images for each person. We set the initial learning rate of the Adam optimizer to  $10^{-3}$  and shrink it by a factor of 0.1 at 80 and 160 epochs until convergence is achieved.

#### 4.3. Comparisons with Traditional Methods

Most traditional person re-identification studies try to learn the most distinctive and stable characteristics to describe the characteristics of each pedestrian. Usually, they use the color feature, texture feature, shape feature, and a fusion of the three features [21]. The comparison results with the traditional method are shown in Table 2.

**Table 2.** The comparison results with traditional methods on Market-1501 dataset.

Methods	Rank-1	mAP
gBiCov [14]	8.28	2.33
LOMO [4]	43.79	22.22
BoW	34.38	14.10
BoW+KISSME [38]	44.4	20.8
ours	95.7	88.1

On the Market-1501 dataset, the experimental results show that the results of our model far exceed the traditional person re-identification methods in both evaluation indicators, where Rank-1 is 95.7% and mAP is 88.1%.

#### 4.4. Comparisons with the State-of-the-Arts

With the advancement of deep learning [8,10,11,24,35,39], person Re-ID has achieved inspiring performance on widely used benchmarks. However, as Table 3 has shown, the parameters of common convolutional neural networks can easily reach millions, such as the GoogleNet used by attention-aware compositional network (ACCN) [10], ResNet50 used by multivariable multi-objective genetic algorithm (MMGA) [11]. Our model has only  $2.4 \times 10^5$  parameters, but it has also reached the state of the art, as shown in Table 4.

**Table 3.** Several classic convolutional neural network model parameters.

Methods	Million Parameters
AlexNet [40]	60
VGG16 [41]	138
GoogleNet [42]	6.8
Inception-v3 [43]	23.2
ResNet50 [44]	25.5
ours	0.24

**Table 4.** Comparison of results on Market-1501 and DukeMTMC-reID.

Methods	Market-1501		DukeMTMC-reID	
	Rank-1	mAP	Rank-1	mAP
ACCN [10]	85.9	66.9	76.8	59.3
MSCAN [39]	83.6	74.3	-	-
MultiScale [3]	88.9	73.1	79.2	60.6
HA-CNN [45]	91.2	75.7	80.5	63.8
AlignedReID [8]	91.0	79.4		
HPM [24]	94.2	82.7	86.4	74.6
MMGA [11]	95.0	87.2	89.5	78.1
UnityStyle+RE [46]	93.2	89.3	85.9	82.3
ours	95.7	88.1	85.3	78.4

On the Market-1501 dataset, our model outperforms most of the state-of-the-art algorithms and exceeds the current best models, i.e., it is attention-driven by +0.7% in Rank-1. Additionally, the proposed model is 1.2% lower than UnityStyle+RE in mAP to achieve the second best result.

On DukeMTMC-reID, our model achieves 85.3% Rank-1 and 78.4% mAP. In Rank-1, our method significantly exceeds most deep learning methods to achieve the second best result in mAP, which is 0.9% lower than that of UnityStyle+RE.

#### 4.5. Ablation Study

The importance of the color information in our proposed hand-crafted feature is self-evident. To verify the effectiveness the role of the relative spatial information (RSI) in the feature and the neural network in the model, we conducted an ablation experiment, and the results experiment are shown in the Table 5.

**Table 5.** The result of ablation experiment.

Index	Methods	Rank-1	mAP
1	without RSI	38.6	18.7
2	with RSI	47.2	25.1
3	without RSI + Att-BLSTM	87.3	73.6
4	with RSI + Att-BLSTM	95.7	88.1

In index-1 and index-2, we only use hand-crafted features, such as in the traditional way, without going through a neural network. Only in index-1, we remove the relative spatial information (RSI) in the feature. Compared with the color feature without RSI, the color feature with RSI has improved Rank-1 and mAP by +8.6% and +7.4%, respectively. Then, in index-3 and index-4, we input the two features into Att-BLSTM separately. As shown in the experimental results after adding Att-BLSTM, whether the input is the color feature with RSI and the color feature without RSI, Rank-1 and mAP are improved a lot. Among them, the effect is the best when the input is the color feature with RSI, and rank-1 and mAP are achieved respectively 95.7% and 88.1%.

## 5. Discussion

By comparing with traditional methods, we think that traditional person re-identification methods pay more attention to the design of features. Although the features designed based on prior knowledge are relatively simple and easy to understand, it is easy to overlook the contextual semantic information, the disadvantages are also obvious, lack of contextual semantic information, and cannot fully express the original image information. However, our model uses the Att-BLSTM to obtain contextual semantic information, and at the same time, due to the attention mechanism, the model can automatically focus on the features that are decisive for the task. Although our model was not optimal in experiments with deep learning methods, the number of parameters in our model was exponentially lower than other methods, which means less loss in computing resources, storage space, and time costs. A small number of parameters means less storage space is required. Therefore, with the popularity of edge computing, a model that combines low-dimensional features and deep learning must play a major role.

The ablation experiment proved the importance of the relative spatial information contained in the color features and the important role of Att-BLSTM in the model. Although the color feature with RSI has been refined based on prior human knowledge, it lacks an in-depth understanding and analysis of features and only stays at the surface layer of the features. The existence of Att-BLSTM enables the model to obtain the contextual relationship contained in the hand-crafted color features, which improves the performance of the model.

## 6. Conclusions

In this paper, we designed a color feature containing relative spatial information, namely the color feature with spatial information. Then, bidirectional long short-term memory (BLSTM) networks with an attention mechanism are used to obtain the contextual relationship contained in the hand-crafted color features. The experiments show that our

proposed method can significantly improve the accuracy of person Re-ID compared with traditional methods. Compared with the current state-of-the-art methods, we use fewer parameters and consume fewer computing resources to achieve the same level as them. In addition, due to the existence of handcraft features, the parameter amount of the model is greatly reduced, which means that the consumption of computing resources and storage resources is greatly reduced. Although we show that our methods outperform others in the Market1501 and DukeMTMC-reID datasets, there is still a large gap between humans and machines. Person changes in clothing, occlusion and other issues will affect the extraction of features, thereby affecting the performance of our proposed model. However, human vision can obtain more high-latitude features and can also use prior knowledge to assist recognition, which are directions for our future research.

**Author Contributions:** Supervision, H.X. and M.B.; Writing—original draft, X.C.; Writing—review & editing, X.C. and Y.L. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Data Availability Statement:** Not Applicable, the study does not report any data.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Bazzani, L.; Cristani, M.; Murino, V. Symmetry-driven accumulation of local features for human characterization and re-identification. *Comput. Vis. Image Underst.* **2013**, *117*, 130–144. [[CrossRef](#)]
2. Zeng, M.; Wu, Z.; Chang, T.; Lei, Z.; Lei, H. Efficient Person Re-identification by Hybrid Spatiogram and Covariance Descriptor. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015.
3. Zhao, R.; Ouyang, W.; Wang, X. Person re-identification by saliency learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *39*, 356–370. [[CrossRef](#)] [[PubMed](#)]
4. Liao, S.; Yang, H.; Zhu, X.; Li, S.Z. Person re-identification by Local Maximal Occurrence representation and metric learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015.
5. Layne, R.; Hospedales, T.M.; Gong, S. Person Re-identification by Attributes. *BMVC* **2012**, *2*, 8.
6. Variator, R.R.; Haloi, M.; Wang, G. Gated Siamese Convolutional Neural Network Architecture for Human Re-Identification. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016; Springer: Cham, Switzerland, 2016.
7. Lin, W.; Shen, C.; Hengel, A. Personnet: Person re-identification with deep convolutional neural networks. *arXiv* **2016**, arXiv:1601.07255.
8. Xuan, Z.; Hao, L.; Xing, F.; Xiang, W.; Jian, S. Alignedreid: Surpassing human-level performance in person re-identification. *arXiv* **2017**, arXiv:1711.08184.
9. Wang, G.A.; Yang, S.; Liu, H.; Wang, Z.; Yang, Y.; Wang, S.; Yu, G.; Zhou, E.; Sun, J. High-Order Information Matters: Learning Relation and Topology for Occluded Person Re-Identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020.
10. Jing, X.; Rui, Z.; Feng, Z.; Wang, H.; Ouyang, W. Attention-aware compositional network for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018.
11. Cai, H.; Wang, Z.; Cheng, J. Multi-Scale Body-Part Mask Guided Attention for Person Re-Identification. In Proceedings of the Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Long Beach, CA, USA, 16–17 June 2019.
12. Leng, Q.; Ye, M.; Tian, Q. A survey of open-world person re-identification. *IEEE Trans. Circuits Syst. Video Technol.* **2019**, *30*, 1092–1108. [[CrossRef](#)]
13. Ye, M.; Shen, J.; Lin, G.; Xiang, T.; Hoi, S. Deep learning for person re-identification: A survey and outlook. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *1*. [[CrossRef](#)] [[PubMed](#)]
14. Ma, B.; Yu, S.; Jurie, F. Covariance descriptor based on bio-inspired features for person re-identification and face verification. *Image Vis. Comput.* **2014**, *32*, 379–390. [[CrossRef](#)]
15. Matsukawa, T.; Okabe, T.; Suzuki, E.; Sato, Y. Hierarchical Gaussian Descriptor for Person Re-identification. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016.
16. Ma, L.; Yang, X.; Tao, D. Person re-identification over camera networks using multi-task distance metric learning. *IEEE Trans. Image Process.* **2014**, *23*, 3656–3670. [[PubMed](#)]
17. Zheng, Z.; Liang, Z.; Yi, Y. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), IEEE Computer Society, Venice, Italy, 22–29 October 2017.
18. Qian, X.; Fu, Y.; Jiang, Y.G.; Xiang, T.; Xue, X. Multi-scale Deep Learning Architectures for Person Re-identification. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017.

19. Swainthand, M.J.; Ballard, D.H. *Indexing Via Color Histograms Object Identification via Histogram*; Springer: Berlin/Heidelberg, Germany, 1990.
20. Hu, H.M.; Fang, W.; Zeng, G.; Hu, Z.; Li, B. A person re-identification algorithm based on pyramid color topology feature. *Multimed. Tools Appl.* **2016**, *76*, 26633–26646. [[CrossRef](#)]
21. Li, Y.; Tian, Z. *Person Re-Identification Based on Color and Texture Feature Fusion*; Springer International Publishing: Berlin/Heidelberg, Germany, 2016.
22. Geng, Y.; Hu, H.M.; Zeng, G.; Zheng, J. A person re-identification algorithm by exploiting region-based feature saliency. *J. Vis. Commun. Image Represent.* **2015**, *29*, 89–102. [[CrossRef](#)]
23. Peng, Z.; Wei, S.; Tian, J.; Qi, Z.; Bo, X. Attention-Based Bidirectional Long Short-Term Memory Networks for Relation Classification. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics 2016, Berlin, Germany, 7–12 August 2016; Volume 2.
24. Fu, Y.; Wei, Y.; Zhou, Y.; Shi, H.; Huang, G.; Wang, X.; Yao, Z.; Huang, T. Horizontal pyramid matching for person re-identification. *Proc. AAAI Conf. Artif. Intell.* **2018**, *33*, 8295–8302. [[CrossRef](#)]
25. Stewart, R.; Andriluka, M.; Ng, A.Y. End-to-end people detection in crowded scenes. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016.
26. Neculoiu, P.; Versteegh, M.; Rotaru, M. Learning Text Similarity with Siamese Recurrent Networks. In Proceedings of the Repl4NLP Workshop at ACL2016, Berlin, Germany, 11 August 2016.
27. Jobson, D.J.; Rahman, Z.U.; Woodell, G.A. A multiscale retinex for bridging the gap between color images and the human observation of scenes. *IEEE Trans. Image Process.* **1997**, *6*, 965–976. [[CrossRef](#)] [[PubMed](#)]
28. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017.
29. Chen, L.C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking atrous convolution for semantic image segmentation. *arXiv* **2017**, arXiv:1706.05587.
30. Guler, R.A.; Trigeorgis, G.; Antonakos, E.; Snape, P.; Kokkinos, I. Densereg: Fully convolutional dense shape regression in-the-wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
31. Xiang, L.; Zheng, W.S.; Wang, X.; Tao, X.; Gong, S. Multi-Scale Learning for Low-Resolution Person Re-Identification. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015.
32. Khan, F.S.; Rao, M.A.; Weijer, J.; Felsberg, M.; Laaksonen, J. Deep Semantic Pyramids for Human Attributes and Action Recognition. In Proceedings of the Scandinavian Conference on Image Analysis, Copenhagen, Denmark, 15–17 June 2015; Springer International Publishing: Berlin/Heidelberg, Germany, 2015.
33. Srivastava, D.; Bakthula, R.; Agarwal, S. Image classification using surf and bag of lbp features constructed by clustering with fixed centers. *Multimed. Tools Appl.* **2018**, *78*, 14129–14153. [[CrossRef](#)]
34. Eitz, M.; Hildebrand, K.; Boubekur, T.; Alexa, M. Sketch-based image retrieval: Benchmark and bag-of-features descriptors. *IEEE Trans. Vis. Comput. Graph.* **2011**, *17*, 1624–1636. [[CrossRef](#)] [[PubMed](#)]
35. Chen, Y.; Zhu, X.; Gong, S. Person Re-identification by Deep Learning Multi-scale Representations. In Proceedings of the 2017 IEEE International Conference on Computer Vision Workshops (ICCVW), Venice, Italy, 22–29 October 2017.
36. Hermans, A.; Beyer, L.; Leibe, B. In defense of the triplet loss for person re-identification. *arXiv* **2017**, arXiv:1703.07737.
37. Zheng, L.; Shen, L.; Lu, T.; Wang, S.; Qi, T. Scalable Person Re-identification: A Benchmark. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015.
38. Kostinger, M.; Hirzer, M.; Wohlhart, P.; Roth, P.M.; Bischof, H. Large scale metric learning from equivalence constraints. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 2288–2295.
39. Song, C.; Yan, H.; Ouyang, W.; Liang, W. Mask-Guided Contrastive Attention Model for Person Re-identification. In Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018.
40. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. *Adv. Neural Inf. Process. Syst.* **2012**, *25*, 2012.
41. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
42. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015.
43. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the Inception Architecture for Computer Vision. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 2818–2826.
44. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016.
45. Li, W.; Zhu, X.; Gong, S. Harmonious Attention Network for Person Re-Identification. In Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018.
46. Liu, C.; Chang, X.; Shen, Y.D. Unity Style Transfer for Person Re-Identification. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020.