Open Access

Journal of Artificial Intelligence & Robotics

Article type: Research Article

AI-enhanced face recognition in cartoons: A multi-domain learning approach

Jianjun Deng[1]; Qinglai Xie[2]; Yao Liu[3]*

[1]Chengdu Aeronautic Polytechnic, Chengdu, Sichuan 610100, China.

[2]College of Physics and Information Engineering, Quanzhou Normal University, Fujian 362000, China.

[3]Quanzhou Bolang Technology Group Co Ltd, Quanzhou, Fujian, China.

***Corresponding Author: Yao Liu**

Quanzhou Bolang Technology Group Co Ltd, Quanzhou, Fujian, China.

Email: yowk0529@gmail.com

Received: Sep 05, 2024

Accepted: Oct 08, 2024

Published Online: Oct 15, 2024

Website: www.joaiar.org

Volume 1 [2024] Issue 1

Abstract

This paper presents our solution to the cartoon photo face recognition competition of the CCF Big Data & Computing Intelligence Contest (CCF BDCI) training track, focusing on symmetry in AI-enhanced recognition. Our approach utilizes a robust baseline of person re-identification, improved by integrating symmetrical data processing techniques. First, we propose a multi-domain learning approach that harmoniously combines face photos and cartoons to train the model. Then, we introduce an identity mining method that symmetrically generates pseudo-labels for part of the test data, outperforming the K-means clustering method. Finally, we test both the original and improved methods under multi-model integration conditions. Results demonstrate that the mAP score of our improved method reached 0.7087, a 58.52% improvement over the original method, securing fourth place in the competition.

Keywords: Person re-identification; Cartoon photo face recognition; Multi-domain learning; Identity mining.

Introduction

CCF Big Data & Computing Intelligence Contest (CCF BDCI) was founded by China Computer Society in 2013. The competition is a large-scale challenge for algorithms, applications, systems and entrepreneurship in the field of big data and artificial intelligence. This paper introduces our solution to the cartoon photo face recognition competition in the training competition.

Face recognition in cartoon photos

In recent years, with the rapid development of computer vision, face recognition has become one of the most important research areas, and more and more researchers have joined in the study of face recognition. Research shows that face recognition has high accuracy in biometric recognition. By analyzing face images, researchers extract useful and discriminative face features to achieve the purpose of face retrieval. At present, with the increasing maturity of face recognition technology, its application field is also expanding. Cartoon face recognition refers to the research field of automatic recognition and analysis of faces in cartoon images. Compared with traditional real face recognition, comic face recognition faces more challenges and difficulties, mainly due to the special nature of comic images and the diversity of artistic creation styles. The research goal of cartoon face recognition is to match and recognize the faces in cartoon images with known faces by using computer vision and pattern recognition technology. This technology has important potential in many applications, in comic analysis, character recognition, face search, comic content automation processing and copyright protection have an important role, such as comic face recognition technology can be used for copyright protection and content monitoring. By identifying and comparing the fa-

Citation: Deng J, Xie Q, Liu Y. AI-enhanced face recognition in cartoons: A multi-domain learning approach. J Artif Intell Robot. 2024; 1(1): 1008.

cial features of the characters in the comic book, it is possible to detect unauthorized copies of the comic book or unauthorized use of the comic book character, thereby protecting the rights of the original author. At present, methods based on machine learning and deep learning are widely used in comic face recognition. These methods include traditional feature matching algorithms, deep learning-based Convolutional Neural Networks (CNN) and generative adversarial networks (Gans). However, there are still the following difficulties in comic face recognition.

Domain differences: There are large domain differences between comic photos and real face photos. Cartoon photos usually have the characteristics of abstraction, unreal and deformation, and the visual characteristics of real face photos are significantly different. This domain difference makes it difficult for the model to accurately capture the key features of the face when processing comic photos, thus affecting the accuracy of face recognition. Incomplete information and distortion: Faces in comic photos are usually affected by artistic styles and drawing techniques, and there may be problems such as incomplete information, blurred details, and deformed shapes. Compared with real face photos, the face features in comic photos may be changed by artistic techniques, such as the scale, proportion and shape of the features may be exaggerated or deformed, making it difficult for the model to accurately match and identify.

Uneven sample: The available data sets for comic photos are generally small and uneven compared to photos of real faces. This may result in the model learning the real face photos more fully during the training process, while the learning of the cartoon photos is relatively insufficient. This results in lower performance of the model on comic photos, making it difficult to perform accurate face recognition.

Pose and expression changes: Characters in comic photographs may be depicted in a variety of different poses and expressions, which causes significant changes in the Appearance and morphology of the human face. Models need to be able to model and recognize different gestures and expressions, which increases the difficulty of face recognition.

Sparse annotation data: Compared with real face photos, the annotation data of cartoon photos is relatively scarce. Due to the special nature of comic photographs and the diversity of artistic creation styles, it is difficult to provide accurate and comprehensive labeling of comic photo datasets. This makes it difficult to obtain enough supervisory signals in the training process, which affects the training and generalization ability of the model.

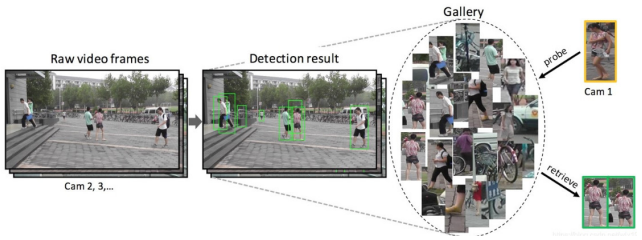


Figure 1: Reidentification algorithm flow.

Re identification algorithm

In recent years, the Re-Identification algorithm (ReID) of deep neural networks has made progress and achieved high performance. The rerecognition algorithm flow is shown in Figure 1. However, many of the most advanced methods design complex network structures and connect multiple branch features. In the literature, some effective training techniques or improvements appear briefly in several papers or source code. This article will personally collect and evaluate these effective training techniques. By adding all the training tips, ResNet50 achieved a 94.5% ranking 1 accuracy and 85.9% mAP on the market [1]. It is worth mentioning that the global features of the type achieve such a further improvement by adjusting the relevant Settings to leverage ResNet101 to extract the global features for the best ReID performance. As shown in Figure 2, a dataset of character photos and cartoons is provided to train the model. There is a large bias between these two different data sets, so how to use the comic data set properly remains a big challenge. Comic photo face recognition is not a standard ReID task, where models are trained and evaluated on data from the same domain [2]. However, ReID technology has made remarkable progress in the field of face recognition, and has proposed some effective feature learning and representation methods. These methods can capture key features in face images, such as color, texture and shape, so as to enhance the robustness and accuracy of comic face recognition algorithms for comic images. At the same time, the domain adaptive method in ReID technology can help solve the domain difference between the cartoon image and the real image.

Figure 2: Photo and comic data section examples.

Through feature alignment and interdomain transfer in different domains, ReID technology can make the model better adapt to the characteristics of comic images and improve the performance of comic face recognition.

Since this task is similar to ReID and ReID can play a good role in the field of comic book face recognition through analysis, we use the strong baseline network BoT-BS in ReID in this

paper [3,4]. Among them, BoT-BS introduced BNNeck to reduce inconsistencies between ID losses (cross-entropy losses) and triplet losses during the training phase. In this paper, we first test the dataset against the original ReID, and then modify the training Settings of the original method, such as the learning rate, optimizer, and loss function, to improve the ReID performance on the new dataset. In addition to modifying BoT-BS, we also focused on how comic data can be used to improve ReID performance on real photo data. Because there is a large bias between photo data and comic data, it is more challenging in face ReID than cross-domain or domain adaptive tasks. We observed that directly merging photos and comics to train models and pre-train models with comic data did not improve ReID performance. Based on the motivation of low-level features, we propose a Multi-Domain Learning (MDL) approach in which the model pre-trains a portion of the photo and comic data and then fine-tune the real photo data [5]. In addition, test sets are allowed. Some unsupervised methods, such as k-means clustering, can generate pseudo-labels for test data [6]. However, due to the poor performance of ReID's model, the pseudo-labels are not accurate enough. Therefore, we propose an Identity Mining (IM) method to generate more accurate pseudo-tags. IM selects a number of samples with different ids with high consistency as the clustering center, and can only select one sample per identity. Then, for each cluster center, tag the ID with the same sample. Unlike k-means clustering, which divides all data into multiple clusters, our IM method simply automatically labels part of the data with high confidence.

While our mAP accuracy was only 12.35% on the original baseline method, we further improved the mAP accuracy to 70.87% through improved methods and multi-model integration. Our contribution can be summarized as follows:

Proposes an MDL strategy that combines real photo and comic data.

An IM method for automatically generating pseudo-labels for part of test data is proposed.

We hit 0.7087 in the mAP score, fourth in the competition.

Related work

We'll cover some of the work of Deep ReID, comic face recognition, and CCF BDCI in this section.

Deep ReID: Deep learning algorithms have been widely used in ReID algorithm. In this section, I will focus on image-based ReID. Image-based ReID, ReID's mainstay, correlates pedestrians through multiple non-overlapping camera views based on a single frame of image information. Image-based ReID has been extensively studied from the two main steps of spatial feature learning and metric learning [7].

To better distinguish the identities of different people, some researchers have explored feature fusion to improve ReID's performance [8]. Proposed a ReID, joint Single Image Representation (SIR) and cross image representation (CIR) learning framework based on Convolutional Neural Networks (CNN), in which SIR And CIR feature representations were jointly optimized to achieve better matching performance [9]. Proposed a deep ReID model. The method combines the output feature mapping of four convolution layers to generate a single feature embedding. The algorithm makes full use of multi-level visual cues to improve ReID performance. However, it requires multiple hand-designed output convolution layers to extract feature representations, making it difficult to learn more discriminative feature representations. In order to further learn more feature representations to distinguish the identities of different people, some researchers explored improving the structure of CNNS to enhance performance, and [10]. Proposed a deep model called Braid Net to learn feature representations of ReID. The method mainly consists of two parts: feature extraction network and feature fusion network [11]. Proposed a deep learning network-based approach to the dislocation problem in ReID. The network uses two stream subnetworks, including a main full image stream (MF-Stream) and a dense semantic directed stream (DSAG-Stream), to extract global features and partial features respectively, and then uses the element-level feature fusion of MF-Stream and DSAG-Stream. Achieve end-to-end dual-stream joint optimization [12].

Proposed a human semantic analysis of ReID. The network has two branches, a normal ReID branch and a semantically split branch, where the semantically split branch generates different semantic parts via pr. For the metric learning step, the core goal is to learn the optimal distance measure, which aims to make feature distances associated with the same person closer than feature distances associated with different people. As a result, measuring loss has become ReID's focus. Various measures of loss have been used to train deep ReID models [13-18]. Used [13] ReID to train the contrast loss of Long Short-Term Memory (LSTM) model. Given a pair of human images as input, this method uses contrast loss function to optimize LSTM network to learn an embedded feature space. If the feature distance is similar, it is closer, while if the feature distance is different, it corresponds to the learning feature space that is far away from each other [14]. Proposed a triplet loss to train a CNN model based on multi-channel segments. With anchor, positive and negative samples as inputs, the method uses triplet loss function to optimize the multi-channel CNN model. In order to learn features, the feature distance belongs to the same person, and the feature distance belongs to the feature space for further learning by different people. However, from the training set to the test set, the generalization of the triplet loss may be weak.

Face recognition in cartoon photos: Traditional face recognition systems already have a large number of face image databases, static, dynamic, different environments, different expressions, different lighting and other circumstances of the face image. However, if the study is the face image required by the individual under special circumstances, the researchers need to collect and organize themselves or shoot and collect themselves. In the case of a large number of samples, the face recognition system mainly includes the following four parts: face image detection, face image preprocessing, face image feature extraction and feature matching. The same is true for comic photo face recognition. In comic photo feature extraction algorithm, researchers have proposed different comic face recognition methods, including algorithms based on traditional feature extraction and machine learning methods, such as the combination of local feature descriptors and classifiers; And deep learning-based methods such as Convolutional Neural Networks (CNN) and generative adversarial networks (Gans). At the same time, in order to promote the research of comic face recognition algorithms, some researchers have constructed comic face data sets. The most famous of these is the Manga109 dataset, which contains 109 comic pages from different comic works for tasks such as comic face detection, recognition, and analysis.

CCF BDCI: CCF BDCI was founded by China Computer Soci-

ety in 2013. The competition is a large-scale challenge for algorithms, applications, systems and entrepreneurship in the field of big data and artificial intelligence. So far, the competition has been successfully held for seven times, and the scale and influence of the competition have been increasing year by year, attracting more than 90,000 participants from more than 1,500 universities, 1,800 enterprises and public institutions and more than 80 scientific research institutions around the world. Among them, the 7th competition in 2019 alone attracted 25,045 teams composed of 28,269 people from 25 countries around the world to participate, and these teams came from 1,282 enterprises such as Google and Tencent, and 1,215 universities such as MIT and Tsinghua University, and submitted more than 80,000 works. The competition has become one of the most influential activities in the field of big data and artificial intelligence in China, and is the first brand of big data comprehensive competition. In 2020, the eighth CCF BDCI Competition will further expand the scale and influence of participation, pay attention to technological development and talent training, and help promote the development of China's big data technology and industrial ecology. In order to help college personnel training, the permanent open training track is specially released to help production, learning and research. Among them, the field of computer vision research is increasingly interested in the recognition and generation of comics. The goal of cartoon recognition research is to study whether a computer can recognize a cartoon from a particular photograph. In recent years, there has been more and more research in this area, and one of the main reasons for this is that it can help understand how humans recognize faces and bridge the gap between human perception and machine recognition. The other main reason is that there are better comic recognition mechanisms, which can be used to synthesize better comics while retaining their intrinsic identity. Past research has suggested that research into human perception of faces in photos and cartoons may help understand how the human brain represents and encodes faces.

Material and methods

In this paper, the original method is used to test the data, and then the method is improved to obtain the best performance.

Baseline Model

The baseline model is important for the final ranking. We used a strong baseline (BoT-BS) [3,4]. proposed by ReID as the baseline model. Figure 3 shows the training strategy and model architecture. The following focuses on the more important parts of the original baseline model.

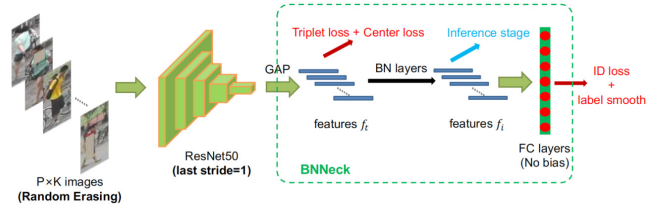


Figure 3: The pipeline of our modififfied baseline.

The final layer of BoT-BS is a fully connected layer that outputs the predictive logic output of the image with a size equal to the number of data set categories n. For an image, we represent y as the real ID and p_i as the logical output of class i. Cross entropy loss is calculated as follows:

$$L(ID) = \sum_{i=1}^{N} -q_i \log(p_i) \begin{cases} q_i = 0, y \neq i \\ q_i = 1, y = i \end{cases} \quad (1)$$

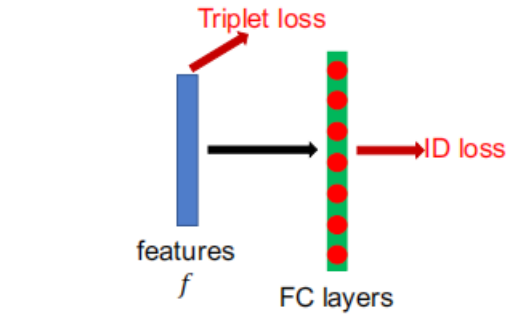
Since the category of the classification is determined by the ID of the cartoon face, this loss function is called ID loss in this paper. However, in person ReID, person IDs of the testing set do not appear in the training set. So it is important to prevent from overfifitting training IDs for the ReID model. A widely used technique to prevent overfifitting for a classifification task is Label Smoothing (LS) proposed in [13]. The construction of q_i is changed to:

$$q_i = \begin{cases} 1 - \frac{N-1}{N}\varepsilon, if\ i = y \\ \varepsilon / N, otherwise \end{cases} \quad (2)$$

Where ε is a constant to encourage the model to be less confident on the training set. In this study, ε is set to be 0.1. When the training set is not large, LS can significantly improve the model performance.

A high spatial resolution always enriches feature granularity PCB [5]. The last spatial down-sampling operation of the backbone network is removed to enlarge the spatial size of the feature map. For convenience, the last spatial down-sampling operation in the backbone network is denoted as the last stride. The last stride equals to 2 for ResNet50 backbone. When fed into an image with 256×128 size, it outputs a feature map with a spatial size of 8×4. If last stride is changed from 2 to 1, then we can obtain a feature map with increased spatial size (16×8). This manipulation only slightly increases the computation cost and does not involve extra training parameters. However, an increased spatial resolution brings significant improvement. The learning rate has great influence on the performance of BoT-BS model. The standard baseline is initially trained at a large and constant learning rate. This choice may result from the expectation that the initial parameters of the model are far from the optimal solution. The higher learning rate can promote the parameter update quickly in the early stage of training and accelerate the convergence of the model. However, using a constant learning rate can cause the training process to be too fast, miss some local optima, or fail to fine-tune parameters when approaching the global optimal. To overcome these problems, a common approach is to use a learning rate decay strategy. Learning rate attenuation allows the size of the learning rate to be gradually reduced in the later stages of training, enabling the model to approach the optimal solution more precisely. Learning rate attenuation can be based on fixed rules, such as attenuation by a fixed step size, or it can be dynamically adjusted according to the performance of the model on the verification set.

A BoT-BS network can map these samples into an embedded vector space by building triples of data, as shown in Figure 4, containing anchor points, positive samples, and negative samples. Then, by comparing the distance between the anchor point and the positive sample with the distance between the anchor point and the negative sample, a measure of similarity, such as Euclidean distance or cosine similarity, can be calculated. By training with BoT-BS networks and triples of data, the representation of embedded vectors can be optimized so that similar samples are closer together in the embedded space and dissimilar samples are further apart. This helps to improve the accuracy and robustness of similarity comparison and matching tasks.



(a) The neck of the standard baseline.

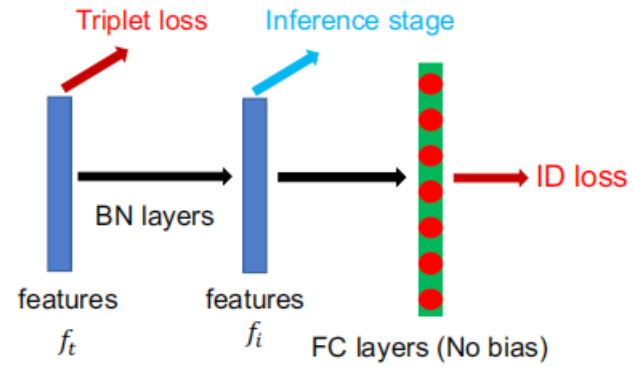


Figure 4: Comparison between standard neck and our designed BN Neck.

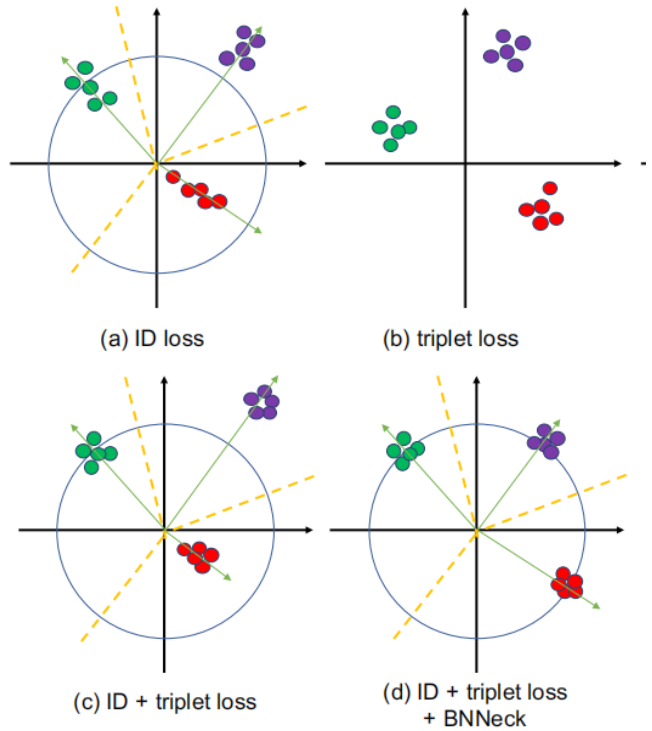


Figure 5: Two-dimensional visualization of sample distribution in the embedding space supervised by **(a)** ID Loss, **(b)** Triplet Loss, **(c)** ID + triplet loss and **(d)** ID + triplet loss + BN Neck. Points of different colors represent embedding features from different classes. The yellow dotted lines stand for the supposed classification hyper planes.

Thus, BoT-BS networks can be used to learn embedding vectors, while triplet data can be used to train and optimize BoT-BS networks so that they can better capture similarity information and distinguish between different samples. They cooperate with each other in the task of similarity learning to improve the performance of the model. Most of works combined ID loss and triplet loss together to train ReID models. As shown in Figure 5a, in the standard baseline, ID loss and triplet loss constrain the same feature f. However, the targets of these two losses are inconsistent in the embedding space. As shown in Figure 6a triplet loss enhances the intra-class compactness and inter class separability in the Euclidean space. Because triplet loss can not provide globally optimal constraint, inter-class distance sometimes is smaller than intra-class distance. A widely used method is to combine ID loss and triplet loss to train the model together. This approach let the model learn more discriminative features. Nevertheless, for image pairs in the embedding space, ID loss mainly optimizes the cosine distances while triplet loss focuses on the Euclidean distances. If we use these two losses to simultaneously optimize a feature vector, their goals may be inconsistent. In the training process, a possible phenomenon is that one loss is reduced, while the other loss is oscillating or even increased.

To overcome the aforementioned problem, we design a structure named as BNNeck shown in Figure 5b. BNNeck only adds a batch normalization (BN) layer after features (and before classifier FC layers). The feature before the BN layer is denoted as ft. We let ft pass through a BN layer to acquire the normalized feature fi. In the training stage, ft and fi are used to compute triplet loss and ID loss, respectively. Normalization balances each dimension of fi. The features are gaussianly distributed near the surface of the hypersphere. This distribution makes the ID loss easier to converge. In addition, BNNeck reduces the constraint of the ID loss on ft. Less constraint from ID loss leads to triplet loss easier to converge at the same time. Thirdly, normalization keeps the compact distribution of features that belong to one same person.

The triplet loss is calculated as:

$$L_{Tri} = [d_p - d_n + \alpha]+, \qquad (3)$$

Where dp and dn are feature distances of positive pair and negative pair. α is the margin of triplet loss, and [z]+ equals to max(z,0). In this paper, α is set to 0.3. However, triplet loss only considers the difference between dp and dn and ignores the absolute values of them. For instance, when dp = 0.3, dn = 0.5, the triplet loss is 0.1. For another case, when dp = 1.3, dn = 1.5, the triplet loss also is 0.1. Triplet loss is determined by two person IDs sampled randomly. It is diffifficult to ensure that dp < dn in the whole training dataset.

$$L_C = \frac{1}{2}\sum_{j=1}^{B} \left\| f_{t_j} - c_{y_j} \right\|_2^2, \qquad (4)$$

Where y_j is the label of the jth image in a mini-batch. c_{y_j} denotes the y_j th class center of deep features. B is the number of batch size. The formulation effectively characterizes the intra-class variations. Minimizing center loss increases intra-class compactness. Our model totally includes three losses as follow:

$$L = L_{ID} + L_{Tri} + \beta L_C \qquad (5)$$

β is the balanced weight of center loss. In our experiments, β is set to be 0.0005.

On the basis of the above model, in order to improve the performance of the data set, we modify some Settings of BOT-BS. The output feature is followed by the BNNeck structure, which separates the ID loss (cross-entropy loss) and triplet loss [21]. Into two distinct embedding Spaces. Triplet losses as soft margin versions are as follows:

$$L_{Triplet} = log[1 + exp(||f_a - f_p||_2^2 - ||f_a - f_n||_2^2 + m)] \quad (6)$$

Hard example mining is used for soft-margin triplet loss. We delete center loss because it does not greatly improve the retrieval performance while increasing computing resources.

At the same time, when the triplet is formed for the triplet loss, when the Anchor is a real person, as shown in Figure 6 (left), both Negative and Positive samples are forced to be cartoons. When the Anchor is a cartoon, as shown in Figure 6 (right), Both Negative and Positive samples are real people. In this way, the triplet is formed to calculate the triplet loss and finally the initial similarity matrix is obtained, and the final matrix is obtained by Rerank operation of the search. And model architecture.

Figure 6: Basic baseline model training strategy and model architecture.

At the same time, in order to obtain more accurate results, the distance matrix generated by the trained model is divided into the query image part and the target image part, and the ranking of the queried image is modified. By comparing the distance between the queried image and the target image, the ranking of the queried image is adjusted when certain conditions are met. Finally, the ranking of the queried image is modified. By comparing the distance between the queried image and the target image, the ranking of the queried image is adjusted when certain conditions are met. The pseudo-code of the algorithm is as follows:

Table 1: Pseudo-code for model joint recognition.

```
dis1 = torch.load(model1)
dis2 = torch.load(model2)
dis3 = torch.load(model3)
g_camids = pickle.load(open('g_camids.pkl','rb'))
distmat = (dis1 + dis2 + dis3) / 3
change = {}
for i in range(len(rank[0])):
 if rank[0][i] > 0.8:
 for j in range(len(query_rank[0][i])):
if query_rank[0][i][j]< 0.7 and
rank[0][query_rank[1][i][j]] < rank[0][i] - 0.1:
 change[i] = query_rank[1][i][j]
 break
rank_list = [g_camids[rank[1][k]] if k not in change else g_camids[rank[1]
[int(change[k])]] for k in range(len(rank[1]))]
```

For better performance, we use the SGD optimizer instead of the Adam optimizer. To improve performance, we trained BoT-BS with deeper backbones and larger sized images.

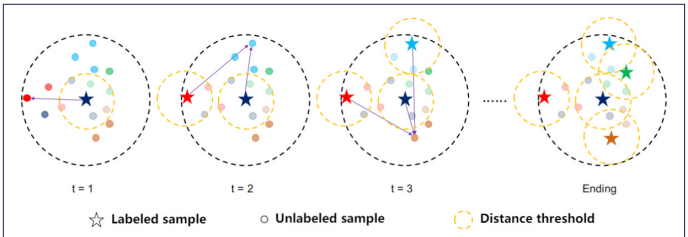


Figure 7: The fifirst stage of our Identity Mining (IM) method. Each step can label a sample with the new ID. The radius of the yellow circle is the distance threshold of the negative pair as dn. The samples outside yellow circles are candidates satisfying the distance constraint. Finally, all samples are covered by at least one yellow circle.

Multi-domain learning

In this section, we will introduce a novel multi-domain learning method to exploit the synthetic data. Both real-world and synthetic data are provided in this challenge, so how to learn discriminative features from two different domains is an important problem. For convenience, the real-world and synthetic data/domains are denoted as DR and DS, respectively. The goal is to train a model on DRUDS and make it achieve better performance on DR. There are two simple solutions as follow.

Solution 1: Directly merging the real-world and synthetic data to train ReID models;

Solution 2: Train a pre-trained model on the synthetic data DS fifirst and then fifine-tune the pre-trained model on the real-world data DR However, these two solutions do not work in the challenge. Since the amount of data in DS is much larger than the amount of data in DR, solution 1 will cause the model to be more biased towards DS. Since there is a large bias between DR and DS, a pre-trained model on DS may not be any better than a pre-trained model on ImageNet. Therefore, Solution 2 is not a good way to solve this problem either. However, some work uses pre-trained models on [23]. Datasets for better performance. This shows that training a pre-trained model on reasonable data is effective. Based on the above discussion, we propose a new MDL method to utilize comic data. The method includes two stages: pre-training stage and fine-tuning stage.

Pre-training phase: The training data for all real photo data is represented as image set r. Then, we construct a new image set s from a part of the cartoon data DS of the random sample. The model pre-trains the new training set rUs to ensure that the pre-trained model is not biased towards DS and the identity of s is not greater than that of r. Specifically, the best performance is when the number of tests is set to 100. We just need to select the first 100 ids of DS.

Fine tuning phase: To further improve the performance of DR, we fine-tuned the pre-trained model without s. Although there is a large domain bias between DR and DS, the low-level features of the two domains are shared. Thus, the first two layers of the pre-trained model are frozen to preserve low-level features during the fine-tuning phase. It is also necessary to reduce the learning rate.

Identity mining

The test set is allowed for unsupervised learning. One widely used method is to use clustering methods to label false labels of

data. Since the test set contains 252 people, we can cluster the test data into 252 categories directly using k-means clustering. However, this method does not work in competition problems because poor models do not give accurate false labels. When we add this auto-annotated data to train the model, we observe that performance gets worse. We believe that it is not necessary to add all the test data to train the model, but it is necessary to ensure the correctness of the false labels. Therefore, we propose an IM method to solve this problem.

The query set is expressed as Q={q_1,q_2,q_3,…,q_m}, and the gallery set is expressed as G={g_1,g_2,g_3,…,g_n}. We use the MDL-trained model to extract the global features of Q and G, expressed as f_Q={f_q1,f_q2,f_q3,…,f_qm } and f_G={f_g1,f_g2,f_g3,…,f_gn } As shown in Figure 9, the first stage is tofind samples with different ids and form a set L={l_1,l_2,l_3…,l_t}. We randomly extract a probe image l_1 to the initial set L:

$$L = \{l_1\}, l_1 \in Q \qquad (7)$$

Then we calculate the distance matrix Dist(Q,L) and de- fifine the distance threshold of the negative pair as dn. The goal is to fifind a sample of new ID to add into set L. To achieve such goal, q_i is considered as a candidate when the sub-matrix $\min(\mathrm{Dist}(q_i, L)) > dn$. However, there may be multiple candidates satisfying this constraint. We select the most dissimilar candidate with all samples in set L as follow:

$$l_t = \mathrm{argmax} \sum \mathrm{Dist}(q_i, L), q_i \in Q \quad (8)$$

$$s.t. \min\left(\mathrm{Dist}(q_i, L)\right) > d_n \qquad (9)$$

Where l_t will be added into the set L with a new ID. We will repeat the process until no qi satisfifies the constraint. After the fifirst step, the set L contains several samples with different IDs. The second step is mining samples belonging to same IDs. Similarly, we defifine the distance threshold of the positive pair as dp. For an anchor image lt, if a sample $x, x \in Q \cup G$ satisfifies $\mathrm{Dist}(x, lt) < dp$, the sample x will be labeled with the same ID with lt. However, x can be labeled with multiple IDs under this constraint. A simple solution is to label x and the most similar lt as thesame ID. Then, these samples are added into set L with pseudo labels. It is noted that only a part of samples are labeled because we set $dp < dn$.

Compared with the k-means clustering, our IM method, which can automatically generate the clustering centers in the fifirst stage, does not need to know the number of classes. However, the proposed method is a local optimization that is sensitive to the initial sample of L. In the future, we will further study it as a global optimization problem. We consider it has the potential to obtain better pseudo labels than other clustering methods.

Experiment

Data set introduction: The data were derived from the Web Caricature dataset, which included a total of 6,042 cartoons and 5,974 photographs belonging to 252 characters. At the same time, since all comic pictures are from the web crawler, the art styles of comics in the data set are diverse.

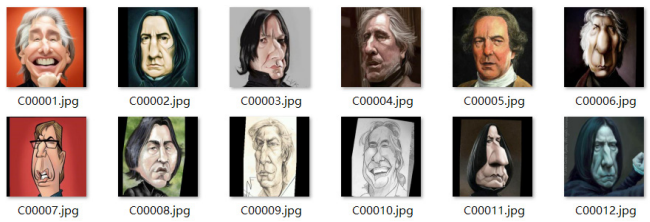


Figure 8: WebCaricature data set 4.2 experimental details.

In addition, the data set does not restrict information such as lighting conditions, posture, expression, occlusion, and age. Given data preprocessing is cutting a good part of the training track data, do not need to face detection and alignment, complete data are available from https://cs.nju.edu.cn/rl/WebCaricature.htm for application.

Experimental details

All images are resized to 320×320. As shown in Figure 3, we used ResNet101 IBN a as the backbone network. For data enhancement, we use random flip, random fill, and random erase. In the training stage, we use soft margin for triplet loss. When forming triplet for triplet loss, when the Anchor is a real person, both Negative sample and Positive sample are forced to be cartoon. When the Anchor is cartoon, we use soft margin for triplet loss. Both Negative and Positive samples are set to real people, resulting in better convergence. Using SGD as the optimizer, the initial learning rate is set to $1e^{-2}$. In addition, we employ a Warmup learning strategy where we train this model with 40 epochs and linearly increase the learning rate from $1e^{-3}$ to $1e^{-2}$.

Figure 9: Base baseline part test results.

Figure 10: We improved the method of partial test results.

In the pre-training phase of MDL, the first 100 identity images were added to the comic face dataset to pre-train the model. These sampled images are fixed at each Epoch. For the functional tuning phase of the MDL, the model is fine-tuned with 40 epoches, where the initial learning rate is set from $1e^{-3}$ for the backbone network layer and from 12 for the full connection layer. For the IM method, the distance threshold d_n is set to 0.49 and d_p is set to 0.23. The features used to calculate the distance are regularized by l_2. At the same time, 130 identities were selected from 252 character categories as L.

Experimental result

Our Web Caricature dataset evaluates the BoT-BS network with our improved approach. Figure 10 shows the test results of the basic baseline. Blue is the sample of correct test results, while red is the sample of wrong test results. In figure 11, part of the test results of our improved method can be seen that our test results are better than those of the basic limit model.

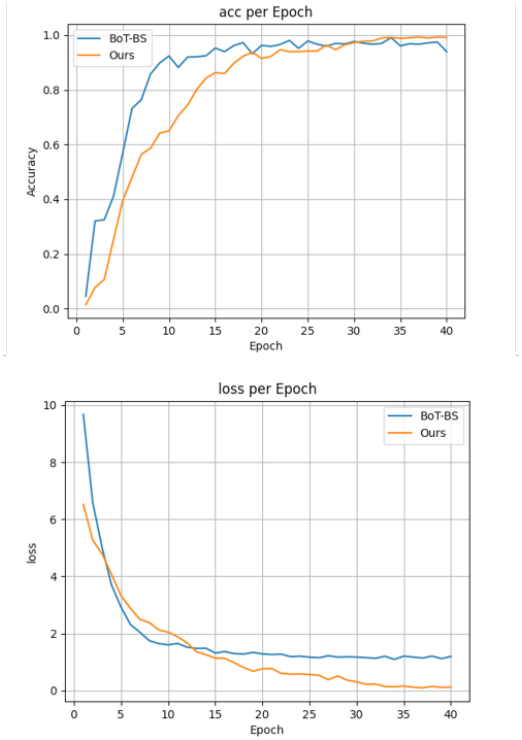


Figure 11: Loss and accuracy of BoT-BS network and our improved method.

Figure 11 are the visual images corresponding to the Accuracy and loss of the BoT-BS network and our improved method in the process of 40 iterations respectively. It can be seen that after 40 iterations, our improved method is better than BoT-BS network in Accuracy and loss.

It can be seen in Table 2 that after 40 rounds of training, the Accuracy of the BoT-BS network and our method in the training set and the mAP in the test set corresponding to them. We can see that our method has a 2.4% higher Accuracy on the training set than BoT-BS. In the test set, mAP was 58.52% higher than BoT-BS.

Table 2: Comparison of methods.

Model	WebCaricature	
	Accuracy	mAP
BoT-BS	0.969	0.1235
Ours	0.993	0.7087

Experimental analysis

It can be seen from the experimental data that BoT-BS performs well in the training set, but the mAP result in the test set is only 12.35%. The reason is that comic photos usually have the characteristics of cartoon, and there are great differences in the painting style, color, texture and shape of different comic photos. Meanwhile, the data set also includes real photos, and these real photos are quite different from comic photos in the above characteristics. The test set may contain comic photos or samples of real people that have not been seen in the training set, and these unknown samples may be more different from the sample of comic photos in the training set. As a result, the BoT-BS network may not accurately capture its features when processing these unknown samples, resulting in a decrease in accuracy. Considering the differences between comic photos and real photos, our proposed multi-domain learning method can expand the training set by using data from multiple fields in the training process, and increase the sample number and diversity of comic photos. In this way, the model can learn the features and attributes of comic photos more fully in the training stage, so as to improve the recognition performance of comic photos. At the same time, the feature representation in different fields (comic photos and real face photos) is learned to better adapt to the comic photos in the test set. By learning shared features between different domains and domain-specific features, the model can better capture the characteristics of comic photos and improve accuracy on the test set. In addition, through the identity mining method, false labels can be automatically generated to incorporate part of the test data into the training process. By generating pseudo-tags for test data and using them as training data, you can expand the training set and increase the diversity and quantity of data. This helps to improve the generalization and adaptability of the model, and can also bring in more information from comic photos and train it with photos of real faces. This helps the model to better understand and distinguish facial features in comic photos, improve the robustness of comic photos, and achieve better performance on the test set.

Results

This paper introduces our improved solution for comic face recognition based on a baseline network (BoT-BS). Because there are great differences between cartoon face and real world face, including painting style, form expression and color use. The BoT-BS network can learn the embedding vector representation that ADAPTS to the features of comic faces, capture the unique features of comic faces, and improve the recognition performance of comic faces. The embedding vector learned by BoT-BS network can be used for similarity comparison and matching of cartoon faces. By calculating the distance or similarity measure between the embedded vectors, the similarity degree between different cartoon faces can be determined, thus supporting the task of cartoon face recognition, face retrieval and face matching. Our research introduces a multi-domain learning strategy that merges real-world and synthetic data to train models effectively. By integrating these data types, the model gains enhanced insight into the distinct features and variations present in comic faces, promoting a richer understanding that accommodates diverse styles and distortions. Our approach inherently appreciates the symmetrical relationship between comic and real faces, enabling the model to achieve improved versatility and generalization. Further, we implement an identity mining technique that strategically generates misleading tags for a segment of the test data. This process amplifies the training dataset by incorporating unlabeled data, which contributes crucial insights that bolster the model's robustness and adaptability. Such methods facilitate a deeper comprehension of the symmetrical features shared by comic and real faces. Ultimately, our method attained a score of 0.7087 in the cartoon photo face recognition competition on Data fountain, securing a commendable 4th place. This result underscores the efficacy and innovation of our model, validating its capability to recognize symmetry across different face types.

Declarations

Acknowledgments: This work is supported by the education and scientific research project for Middle-Aged and Young Teachers of Fujian Province (No. JAT210314).

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

Data Availability: The data used in this study were obtained from https://www.datafountain.cn/competitions/483/datasets. The dataset can be downloaded from the aforementioned website.

References

1. Zheng L, Shen L, Tian L, Wang S, Wang J, et al. Scalable Person Re-Identification: A Benchmark. In Proceedings of the IEEE international conference on computer vision. 2015; 1116-1124.
2. Dietlmeier J, Antony J, McGuinness K, O'Connor N E. How Important Are Faces for Person Re-Identification? In 2020 25th International Conference on Pattern Recognition (ICPR); IEEE. 2021; 6912-6919.
3. Luo H, Gu Y, Liao X, Lai S, Jiang W. Bag of Tricks and a Strong Baseline for Deep Person Re-Identification. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. 2019; 0-0.
4. Luo H, Jiang W, Gu Y, Liu F, Liao X, et al. A Strong Baseline and Batch Normalization Neck for Deep Person Re-Identification. IEEE Transactions on Multimedia. 2019; 22(10): 2597-2609.
5. He S, Luo H, Chen W, Zhang M, Zhang Y, et al. Multi-Domain Learning and Identity Mining for Vehicle Re-Identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. 2020; 582-583.
6. Tatarian K, Couceiro M S, Ribeiro E P, Faria D R. Stepping-Stones to Transhumanism: An Emg-Controlled Low-Cost Prosthetic Hand for Academia. In 2018 International Conference on Intelligent Systems (IS); IEEE. 2018; 807-812.
7. Hou L, Liu Q, Zi Y, Zhou Y, Zhai R. State-of-the-Art Deep Person Re-Identification: A Review. In 2020 7th International Conference on Information Science and Control Engineering (ICISCE); IEEE. 2020; 1328-1334.
8. Wang F, Zuo W, Lin L, Zhang D, Zhang L. Joint Learning of Single-Image and Cross-Image Representations for Person Re-Identification. In Proceedings of the IEEE conference on computer vision and pattern recognition. 2016; 1288-1296.
9. Wang Y, Wang L, You Y, Zou X, Chen V, et al. Resource Aware Person Re-Identification across Multiple Resolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition. 2018; 8042-8051.
10. Wang Y, Chen Z, Wu F, Wang G. Person Re-Identification with Cascaded Pairwise Convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018; 1470-1478.
11. Zhang Z, Lan C, Zeng W, Chen Z. Densely Semantically Aligned Person Re-Identification. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019; 667-676.
12. Kalayeh M M, Basaran E, Gökmen M, Kamasak ME, Shah M. Human Semantic Parsing for Person Re-Identification. In Proceedings of the IEEE conference on computer vision and pattern recognition. 2018; 1062-1071.
13. Varior R R, Shuai B, Lu J, Xu D, Wang G. A Siamese Long Short-Term Memory Architecture for Human Re-Identification. In Computer Vision-ECCV 2016: 14th European Conference, Amsterdam, the Netherlands, October 11-14, 2016, Proceedings, Part VII 14; Springer. 2016; 135-153.
14. Cheng D, Gong Y, Zhou S, Wang J, Zheng N. Person Re-Identification by Multi-Channel Parts-Based Cnn with Improved Triplet Loss Function. In Proceedings of the iEEE conference on computer vision and pattern recognition. 2016; 1335-1344.
15. Chen W, Chen X, Zhang J, Huang K. Beyond Triplet Loss: A Deep Quadruplet Network for Person Re-Identification. In Proceedings of the IEEE conference on computer vision and pattern recognition. 2017; 403-412.
16. Zhou S, Wang J, Wang J, Gong Y, Zheng N. Point to Set Similarity Based Deep Feature Learning for Person Re-Identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017; 3741-3750.
17. Yu R, Dou Z, Bai S, Zhang Z, Xu Y, et al. Hard-Aware Point-to-Set Deep Metric for Person Re-Identification. In Proceedings of the European Conference on Computer Vision (ECCV). 2018; 188-204.
18. Chen D, Xu D, Li H, Sebe N, Wang X. Group Consistent Similarity Learning via Deep Crf for Person Re-Identification. In Proceedings of the IEEE conference on computer vision and pattern recognition. 2018; 8649-8658.
19. Zhang M, Cheng Q, Luo F, Ye L. A Triplet Nonlocal Neural Network with Dual-Anchor Triplet Loss for High-Resolution Remote Sensing Image Retrieval. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. 2021; 14: 2711-2723.
20. Shen Y, Xiao T, Li H, Yi S, Wang X. Learning Deep Neural Networks for Vehicle Re-Id with Visual-Spatio-Temporal Path Proposals. In Proceedings of the IEEE international conference on computer vision. 2017; 1900-1909.
21. Hermans A, Beyer L, Leibe B. In Defense of the Triplet Loss for Person Re-Identification. arXiv preprint. 2017. arXiv:1703.077372017.
22. Li Q, Yuan Y, Wang Q. Hyperspectral Image Super-Resolution via Multi-Domain Feature Learning. Neurocomputing. 2022; 472: 85-94.
23. Yang L, Luo, P, Change Loy, C, Tang, X. A Large-Scale Car Dataset for Fine-Grained Categorization and Verification. In Proceedings of the IEEE conference on computer vision and pattern recognition. 2015; 3973-3981.