Cross-Domain Attention and Center Loss for Sketch Re-Identification

Fengyao Zhu, Yu Zhu, *Member, IEEE*, Xiaoben Jiang, and Jiongyao Ye

***Abstract*— Matching all RGB photos of the target person in the gallery database with the full-body sketch image drawn by the professional is defined as Sketch re-identification (Sketch Re-id). The big gap between the sketch domain and RGB domain makes Sketch Re-id challenging. This paper addresses the problem by proposing a new framework to obtain domain-invariant features, which uses CNN as the backbone. To make the model focus more on the regions related to the sketch image in the RGB photo, we propose a novel cross-domain attention (CDA) mechanism. It uses different ways of splitting feature maps in its two branches and calculates the relationship between different parts in the sketch images and RGB photos. Moreover, we designed the cross-domain center loss (CDC), which breaks through the limitations that datasets need to be in the same domain in the traditional center loss. It effectively reduces the gap between two domains and makes the features with the same ID closer. The experiment is performed on the Sketch Re-id dataset. Each person has one sketch image and two RGB photos. To evaluate the generalization, we also experimented on two popular sketch-photo face datasets. The result in the Sketch Re-id dataset shows the model performs 3.7% higher than the previous methods. And the result in the CUHK student dataset performs 0.38% higher than the state-of-the-art methods.**

***Index Terms*— Sketch re-identification, cross-domain attention, domain-invariant feature, center loss.**

I. Introduction

THE person re-identification (Re-id) has aroused special attention with the greater development on multi-target detection and multi-target tracking [1], [2]. Common Re-id task includes query images and a gallery consisting of photos captured by various cameras. The query images and the gallery are always in the same domain. And the main goal of the Re-id task is to find the right match in the gallery for each query image. But the query image in the real world is not always in the RGB domain. In real crimes, convicts tend to appear in a place without cameras. Therefore, the RGB photo of the convict is difficult to get. Police usually draw a sketch image of the convict from the description by the witnesses. The sketch image contains many appearance characteristics of the convict. It can be found that the query image is invariably a sketch image but not an RGB photo. This situation has also been confirmed by criminal experts and law enforcement, which was often neglected in past research. Face recognition research [3] and [4] have studied how to match the facial sketch images with the facial photos dataset, in which the sketch images are obtained by professional painters from the description of witnesses.

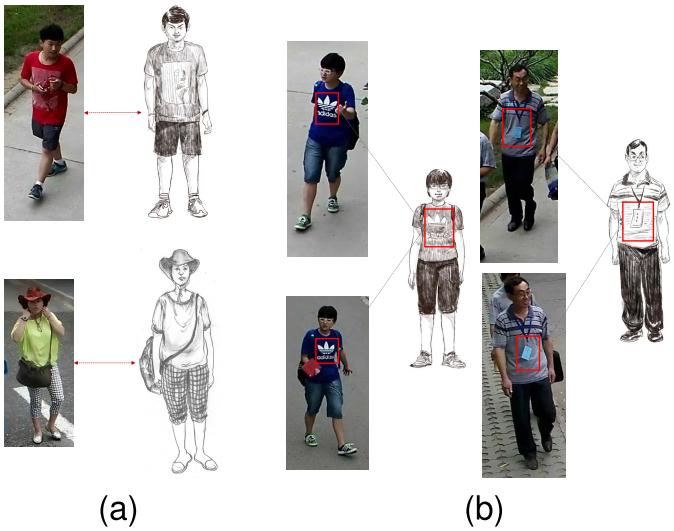

Fig. 1. Some sketch-photo sample pairs in Sketch Re-id dataset. (a) sketch images often lack a lot of color information and RGB photos have rich color information and postures. (b) sketch images and RGB photos have similar local information, like the logo and badges.

In this paper, we define this sketch pedestrian re-identification task as Sketch Re-id, in which the query image is a sketch image and the gallery database is RGB photos. Due to the sketch domain and RGB domain, Sketch Re-id can be regarded as a sketch-based cross-domain recognition task. Meanwhile, the big gaps between two different domains make this task well challenging. As shown in Fig. 1 (a), on account of different painting styles and lack of rich color information, sketch images are extremely abstract. But RGB photos captured by different cameras have rich color information and body postures.

Global and local features are widely used in person Re-id tasks. The global feature contains the global information of one picture. And the local feature contains the information

Manuscript received 26 December 2021; revised 28 June 2022 and 2 September 2022; accepted 12 September 2022. Date of publication 22 September 2022; date of current version 3 October 2022. This work was supported in part by the Natural Science Foundation of Shanghai under Grant 19ZR1413400, in part by the National Natural Science Foundation of China under Grant 82170110, and in part by the Science and Technology Commission of Shanghai Municipality under Grant 20DZ2254400. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Emanuele Maiorana. *(Corresponding author: Yu Zhu.)*

Fengyao Zhu, Xiaoben Jiang, and Jiongyao Ye are with the School of Electronics and Communication Engineering, East China University of Science and Technology, Shanghai 200237, China.

Yu Zhu is with the School of Electronics and Communication Engineering, East China University of Science and Technology, Shanghai 200237, China, and also with the Shanghai Engineer and Technology Research Center of Internet of Things for Respiratory Medicine, Shanghai 200032, China (e-mail: zhuyu@ecust.edu.cn).

Digital Object Identifier 10.1109/TIFS.2022.3208811

of a certain region in the picture. Some similar local features exist in both sketch images and RGB photos, like the logo and badges in Fig. 1 (b). The model needs to extract these special cross-domain local features. We explore the traditional Transformer [5] and multi-modal Transformers [6], [7], [8] to address the cross-domain limitations. The traditional Transformer [5] plays an important role by using cross-attention. In the mid-sub-layer of the traditional Transformer [5] decoder, the query is the output of the previous decoder, and the key and value are the output of the encoder. Differently, the query, key, and value in multi-modal Transformers [6], [7], [8] come from two domains. This makes multi-modal Transformers [6], [7], [8] catch the features and relationships between different domains. Besides, attention mechanisms are popular in recent years [9]. Attention mechanisms assign different weights to different areas in the feature map, which means that not all areas are the same important. The model should focus on the useful areas in the feature map a lot to perform better. In general, the important areas occupy higher weight while other areas are assigned lower weight. The attention block makes the model focus on the most informative area.

We designed the cross-domain attention (CDA) mechanism to capture similar regions related to the sketch image in the RGB photo. It includes a global branch and a local branch. The difference between the two branches is the way of splitting the feature map. In the global branch, we use the way that splitting the feature map by each row and column to get patches, which calculates the relationship between each patch in the sketch feature map and RGB feature map. But the local branch applies the way of splitting the feature map by each row. Each row maps a part in the input image (e.g., a person's head, chest, or legs). The relationship between rows in two domains' feature maps of the same person is calculated in this branch, which is also the relationship between different parts of the input images in two domains. Through the guidance of the sketch feature map, the whole attention block can give different weights to the areas in the RGB feature map. We also propose the cross-domain center (CDC) loss to reduce the gap between two domains. Different from the common person Re-id dataset with a large number of images, the collection of sketch dataset [10] is difficult. There are 200 sketch images painted by 5 painters and 400 RGB photos captured by two cameras in the Sketch Re-id dataset.

The main contributions of this paper are as follows:

(1) A novel cross-domain attention (CDA) mechanism is applied to address the Sketch Re-id task, which contains two branches that use different ways to get patches. The global branch uses the way that splitting the feature maps by each row and column. And the local branch applies the way of splitting the feature maps by each row.

(2) To reduce the gap between two domains and make the features with the same ID closer, we propose the newly cross-domain center (CDC) loss. It maps the features of different sets of domains into a common space, which breaks through the limitation of the traditional center loss.

(3) The accuracy of the Sketch Re-id dataset [10] has increased by 3.7% (Rank-1) and 1.0% (Rank-5) through adding the proposed CDA and CDC. We also test on other sketch-photo face datasets (CUHK student dataset [4] and CUFSF [11]) to test the generalization of the model. Compared with the state-of-the-art methods, our result shows that the rank-1 achieves a 0.38% increase in the CUHK student dataset and a close performance in CUFSF.

II. Related Work

A. Person Re-Identification

Person re-identification mainly solved the problem of pedestrian recognition [12], [13]. The solutions to the problem can be divided into representation learning and metric learning. Representation learning did not consider the similarity between the same ID images but directly regraded the Re-id problem as a classification problem [14], [15] Through marking attributes for the pedestrians, Lin *et al.* [16] improved the generalization ability of the model. Only regarding the Re-id task as a classification problem alone cannot focus on the features of the same ID images. More and more research pay attention to metric learning in recent years. Meanwhile, a variety of loss functions were proposed, like triplet loss [17], [18] and pair verification loss [19]. We usually define the two images of the same ID as a positive pair and two images of different ID as a negative pair. The goal of metric loss is to minimize the distance between the positive pair and increase the distance between the negative pair.

B. Part Features Deep Learning

There has been lot of research to learn a global feature to represent the images of one person. Yet if we only consider the global features but ignore the local features, the model will lose much detailed spatial information. Many works consider the way of dividing images into parts to capture the local features. Compared with the traditional global classification loss in the representation learning, [20] put forward a novel part loss, which automatically detected human body parts and computed the classification loss of each part separately. Aligned-Reid [12] extracted a global feature, which was jointly learned with local features. Sun *et al.* [21] employed the part-level features that offered fine-grained information for pedestrian image description and just considered content consistency within each part of images for precise part location.

C. Visual Transformer

Transformer [5] had shown impressive performance for natural language processing tasks because of the strong ability in modeling long-range context information. Recently, some efficient architectures based on the Transformer have been used for a variety of computer vision tasks. Particularly, Vision Transformer [22] (ViT) reshaped the image into a sequence of flattened patches and input them to the transformer encoder for image classification. Swin Transformer [23] proposed a hierarchical Transformer whose representation is computed with shifted window. It has the flexibility to model at various scales and has linear computational complexity concerning image size. DETR [24] used a common CNN to extract the image's global features, which were taken as the input to a

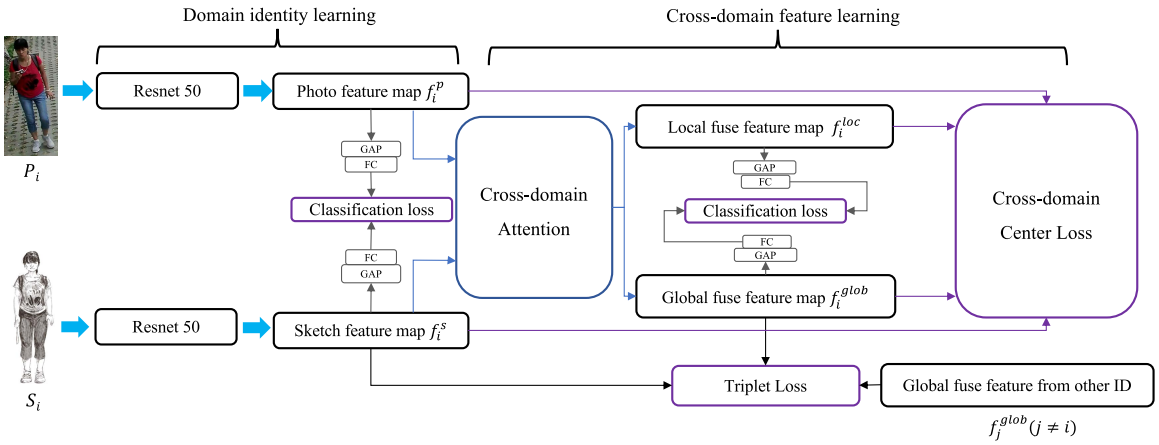

Fig. 2. Proposed network consists of the domain identity learning part and cross-domain feature learning part. Resnet50 is used as the backbone in the first part. The blue block and purple block is the CDA and CDC loss in the second part.

transformer-based encoder-decoder architecture. M2TR [25] adopted a multi-scale transformer that integrates multi-scale information for Deepfake detection.

D. Attention Mechanisms

Adding attention modules to the CNN network has become an effective method to improve recognition accuracy. Many attention mechanisms are designed based on spatial information or channels. Reference [26] put forward a parameter-free spatial attention layer for Person Re-Identification. Hu *et al.* [27] considered the influence of the interdependence between different channels of the feature map and designed the SE attention module. ECA [28] was based on the SE module, which uses the faster adaptive 1D convolution to calculate the attention weights of the channels. The combination of spatial information and channels also performs better. Reference [29] got the attention maps from the channels and spatial information and came up with a simple convolutional attention module. DCA [30] connected the adjacent attention module to make the information flow among attention modules. Self-attention is a highly efficient attention mechanism too. Dual attention network [31] modeled the semantic interdependencies based on the self-attention mechanism, which contained the spatial and channel information.

E. Face Sketch Recognition

Face sketch recognition is one of the most important tasks in the area of heterogeneous face recognition. It refers to face recognition across sketch and RGB domains. There are much research on face sketch recognition in recent years. Lots of methods attempt to use the latent subspace to capture the domain invariant features. G-HFR [32] used the graphical representation and the Markov network to consider the spatial compatibility between neighboring image patches. It also used the designed CRSM to measure the similarity between two different graphical representations. Wang and Tang [11] integrated sparse feature selection with support vector regression and adopted Markov Random Fields. Lin and Tang [33] proposed the Common Discriminant Feature Extraction algorithm, which transformed the different domain samples into a common feature space. Huo *et al.* [34] developed a minibatch proximal point algorithm to make efficient optimization. MvDA [35] seeks a common discriminant space by jointly learning multiple view-specific linear transforms. Peng *et al.* [36] proposes an algorithm that exploits the semantic information integrated with deep convolutional neural networks to fully exploit the identical semantic clue among cross-domain face images. CDFL [37] directly learned the discriminative features from raw pixels for face representation. Besides, some models also use the way of image synthesis to reduce the gap between the different domains [38], which can efficiently improve the recognition performance. CFSS [39] was composed of a multiple feature generator and a cascaded low-rank representation. IA-CycleGAN applies the perceptual loss to supervise the image generation network, which pays more attention to the key facial regions. Luo *et al.* [40] uses the memory module to explore the prototypical style patterns of the reference domain.

III. Approach

A. Network Structure

The proposed model consists of the domain identity learning part and the cross-domain feature learning part. The identity learning part extracts the specific features of sketch images and RGB photos. And the cross-domain feature learning part can reduce the gap between the sketch and RGB through our designed CDA and CDC loss. The training dataset D^{train} in this paper is defined as $\left\{P_i^{1\sim N}, S_i^{1\sim M}\right\}_{i=1\sim K}$, which contains K labels with N RGB photos and M sketch images. Specifically, the domain identity learning part has two branches. The backbone of each branch is the Resnet50, which is popular in retrieval tasks. Each backbone gets a sketch image S_i or an RGB P_i photo as the input during the training stage. Then the output sketch feature map f_i^s and RGB feature map f_i^p from the last convolution layer ($C \times H \times W$, the channel

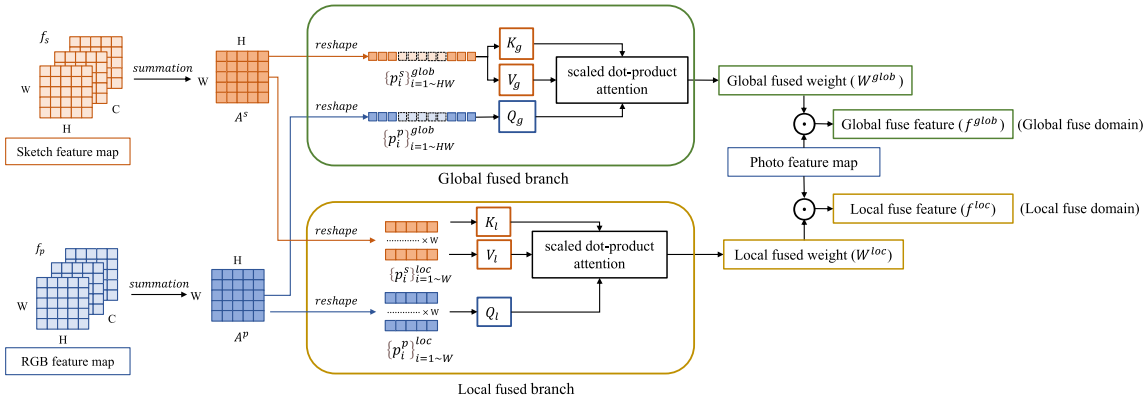

Fig. 3. Cross-domain attention mechanism has a global fused branch and a local fused branch. Each branch calculates different fused weight.

number of C is 2048, and the spatial size $H \times W$ is 7×7) of each Resnet50 are used in the cross-domain feature learning part. Because of the limitation of the backbone, the domain identity learning part can only distinguish different people in a specific domain. The cross-domain feature learning part is designed to solve this problem, which uses the cross-domain attention(CDA) mechanism and cross-domain center loss. The CDA has two branches to get weight and output fused feature maps. As shown in Fig.2., the f_i^{glob} and f_i^{loc} are named as the global fused feature map and local fused feature map. The cross-domain center loss can lower the gap between f_i^s, f_i^p, f_i^{glob} and f_i^{loc} with optimizing the common space center of each label. Common classification loss and triplet loss are also applied to learn the identity information of each person. Each feature map is used for calculating the classification loss. The triplet set in triplet loss consists of an anchor f_i^s, a positive sample f_i^{glob}, and a negative sample f_j^{glob}. The f_i^s, f_i^{glob} have the same label and the f_j^{glob} has a different label. In this way, the triplet set can be written as $\left\{f_i^s, f_i^{glob}, f_j^{glob}\right\}$. The f^s and f^{glob} are used in the validation phase. The purpose of our model is to calculate the similarity between the given query sketch image and each RGB photo in the gallery, which determines the final retrieval result.

B. Cross-Domain Attention Mechanism

The details of the cross-domain attention mechanism are shown in Fig. 3. Not all regions in RGB photos are equally important. Some regions related to the sketch images need to be focused on more. Referring to the mid-sub-layer of traditional Transformer decoder [5], we design the cross-domain attention mechanism, which has a global fused branch and a local fused branch. Both branches apply the scaled dot-product attention mechanism, which needs a query, a key, and a value. The query is got from the RGB feature map, but the key and value are got from the sketch feature map. The result of the scaled dot-product attention mechanism calculation is the cross-domain fused weight. The main difference between the two branches is the way of splitting feature maps. In general, the goal of this attention is to capture the relevant areas in RGB photos through sketch images. As a result, the model can focus on similar and important local areas in sketch images and RGB photos.

Given the feature maps, the first operation is a summation for each pixel along the channels. Two 2-D matrixes A^s and A^p with shape $H \times W$ can be got after the summation. These matrixes indicate the importance of different positions, which are also used as the inputs of the next global fused branch and local fused branch.

1) Global Fused Branch: The splitting way like the ViT [22] is adopted in this branch: The matrix A^s and A^p are divided into HW patches $\left\{p_i^s\right\}_{i=1\sim HW}^{glob}\left(p_i^s \in R^{1\times 1}\right)$ and $\left\{p_i^p\right\}_{i=1\sim HW}^{glob}\left(p_i^p \in R^{1\times 1}\right)$. The $\left\{p_i^s\right\}_{i=1\sim HW}^{glob}$ from the sketch domain are passed into 2 separate linear layers to obtain the global key K_g and global value V_g. The $\left\{p_i^p\right\}_{i=1\sim HW}^{glob}$ from the RGB domain are passed into another linear layer to obtain global query Q_g. With Q_g, K_g and V_g, the global fused weight W^{glob} can be calculated through a scaled dot-product attention mechanism. To be specific, the dot product of Q_g is computed with the K_g. Then we use a Softmax function to obtain the weight summation over V_g. The W^{glob} can be written as equation (1):

$$W^{glob} = \text{softmax}\left(Q_g \cdot K_g^T\right) \cdot V_g \tag{1}$$

It can be found that the W^{glob} fuses the global information from the sketch domain and the RGB domain. Finally, the output global fused feature map f^{glob} is the Hadamard product of the global fused weight W^{glob} and f^p:

$$f^{glob} = W^{glob} \odot f^p \tag{2}$$

2) Local Fused Branch: A new splitting way is designed in the local fused branch. This way is based on the relationship between the input image and the feature map. In the process of extracting picture features by the model, the size of the feature map is gradually cut down with the number of channels increasing. Therefore, Each region in the feature map is mapped into a different area in the input picture. And the rows in f^p or f^s are mapped into the local areas in the sketch image or the RGB photo (e.g., a pedestrian's head, chest, or leg). Due to this, the new splitting way is as follows: the matrix A^s and A^p are divided into W patches $\left\{p_i^s\right\}_{i=1\sim W}^{loc}\left(p_i^s \in R^{1\times H}\right)$

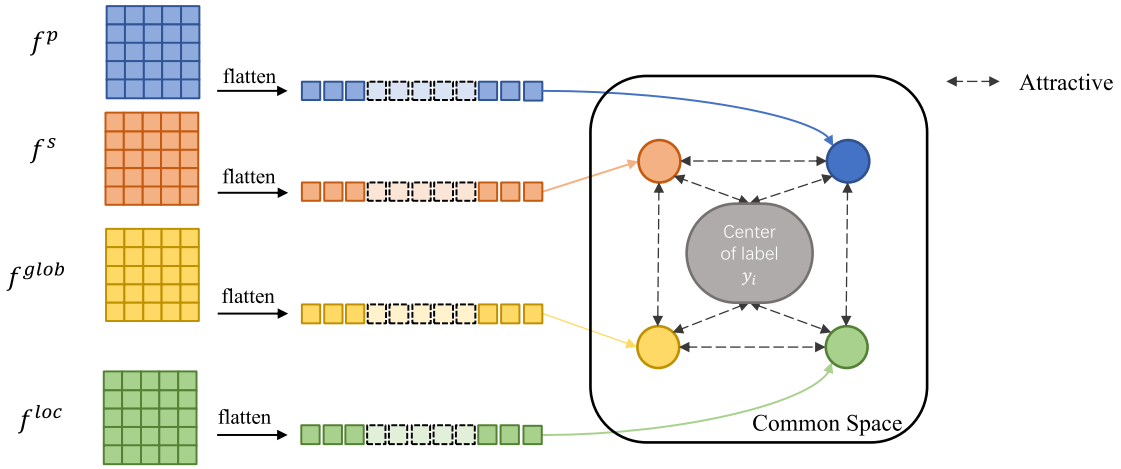

Fig. 4. Illustration of the cross-domain center loss. The distance of features belonging to the same label across all domains can be minimized by optimizing the common space center.

and $\{p_i^p\}_{i=1\sim W}^{loc}$ $(p_i^p \in R^{1\times H})$. Each patch is the rows in the A^s or A^p. In this way, the relationship of different domains' body parts can be calculated. Same as the global fused branch, the $\{p_i^s\}_{i=1\sim W}^{loc}$ from the sketch domain are first passed into 2 separate linear layers to obtain the local key K_l and local value V_l. The $\{p_i^p\}_{i=1\sim W}^{loc}$ from the RGB domain are passed into another linear layer to obtain local query Q_l. Through the scaled dot-product attention mechanism, the final W^{loc} can be calculated by the Q_l, K_l, and V_l:

$$W^{loc} = \text{softmax}\left(Q_l \cdot K_l^T\right) \cdot V_l \tag{3}$$

The local fused weight W^{loc} fuses the local information from the sketch domain and RGB domain, which is different from the global fused branch. The final output local fused feature map f^{loc} can be got from the Hadamard product of the local fused weight W^{loc} and f^p:

$$f^{loc} = W^{loc} \odot f^p \tag{4}$$

3) New Fused Domains: Since the information of sketch domain and RGB domain has fusion and interaction in different ways, this paper defines a global fused domain and a local fused domain. The f^{glob} is regarded as the feature map of the global fused domain. And f^{loc} is regarded as the feature map of the local fused domain.

C. Loss Functions

Loss function plays an important role in the model training process. The classification loss L_{cls} and triplet loss L_{tri} can learn the identify features. Besides, a novel cross-domain center loss L_{cdc} is proposed. The L_{cdc} can help with learning the domain invariant features. overall, this paper uses three types of loss functions. The model can be more robust by jointly using three loss functions:

$$L_{total} = \alpha L_{tri} + \beta L_{cdc} + L_{cls} \tag{5}$$

where α, β are the hyper-parameters designed for L_{tri} and L_{cdc}. By default, we set the $\alpha = 10$ and $\beta = 0.001$. This joint loss function in Equation (5). is optimized by the Adaptive Moment Estimation. The details are shown in the experiment setting.

1) Cross-Domain Center Loss: The traditional center loss [41] was widely used in face retrieval tasks. It strengthens the model's ability to extract features in pictures through the center point changing. In the Sketch Re-id task, the features are from the sketch domain or RGB domain. But the limitation of the past center loss is the pictures in the dataset must be in the same domain. This causes the Sketch Re-id cannot use the traditional center loss. So we propose a new cross-domain center loss based on the traditional center loss [41], [42] to solve Sketch Re-id. Details of this loss are shown in Fig. 4. There is no domain limitation in the cross-domain center loss. The calculation process is as follow: there are N pedestrian labels and d domains in the given feature set $\{f_i^d\}_{i=1}^N (d \in \{s, p, glob, loc\})$($s$ is the sketch domain, p is the RGB domain, $glob$ and loc are the global fused domain and local fused domain). with the $\{f_i^d\}_{i=1}^N$, the cross-domain center loss can be written as Equation (6):

$$L_{cdc} = \frac{1}{2}\sum_{i=1}^{N}\sum_{d\in\{s,p,glob,loc\}}\left\|v_i^d - C_i\right\|_2^2 \tag{6}$$

the C_i is the common space center of pedestrian label i and v is the vector of a specific domain. The model can minimize the distance between the different domain features and their corresponding centers within each training batch through the cross-domain center loss. Each label center C_j will be updated by the features of label j in all domains after each training iteration. The ΔC_j is as Equation (7) shown. $\delta(i, j)$ is a conditional judgment function that judges whether the labels are the same. With updating the center C_j, the distance between the same label's sketch features and RGB features in a batch can be shortened.

$$\Delta C_j = \frac{\sum_{i=1}^{N}\sum_{d\in\{s,p,glob,loc\}}\delta(i, j)\left(C_j - v_j^d\right)}{1 + \sum_{i=1}^{N}\delta(i, j)} \tag{7}$$

TABLE I

NUMBER OF SKETCH IMAGE LABELS FOR EACH PAINTING STYLE, TRAINING, AND TESTING ON THE SKETCH RE-ID DATASET

Style	Numbers	Training	Testing
A	46	34	12
B	20	15	5
C	79	60	19
D	33	25	8
E	22	16	6
Total	200	150	50

$$\delta(i, j) = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases} \tag{8}$$

2) Classification Loss: As shown in Fig. 2, four feature maps are passed through the global average pool to obtain four feature vectors. They are finally used to calculate the classification loss and predict their label. The same label's identity information in the sketch domain and RGB domain can be learned through the classification loss:

$$L_{cls} = \frac{1}{N} \sum_{k=1}^{N} -\log\left(\frac{\exp\left(f_{y_k}\right)}{\sum_j \exp\left(f_j\right)}\right) \tag{9}$$

f_j is the jth element of the predict score that comes from the last FC-layer. y_k is the ground-truth ID of pedestrian sample K.

3) Triplet Loss: The triplet loss is a common metric learning method, which is widely used in pedestrian re-id tasks. Motivated by the margin, triplet loss can make the positive pair close and the negative pair away [17], [43]. The goal of triplet loss in this paper is to minimize the distance between the same label sketch image and RGB photo and increase the distance between different label sketch images and RGB photos. The triplet loss is in the form of:

$$L_{tri} = \sum_{\{f_i^s, f_i^{glob}, f_j^{glob}\}} \max\left(D_{a,p} - D_{a,n} + m, 0\right) \tag{10}$$

$$D_{a,p} = \left\| F(f_i^s) - F(f_i^{glob}) \right\|_2^2 \tag{11}$$

$$D_{a,n} = \left\| F(f_i^s) - F(f_j^{glob}) \right\|_2^2 \tag{12}$$

f_i^{glob} and f_i^s are the global fused feature map and sketch feature map of ID i, which are the most dissimilar pair with same ID. f_j^{glob} and f_i^s are the global fused feature map and sketch feature map of ID j and i, which are the most similar pair with a different ID. The margin m is set as 0.3. The function $F(\cdot)$ means flattening the feature map to a 1-d vector. The L2 distances are defined as the similarity between different features.

IV. EXPERIMENT

A. Dataset

1) Sketch Re-Id Dataset: There is one public Sketch Re-id dataset proposed by [10] for the cross-domain Re-id task, whose sketch images are whole-body of pedestrians. This dataset contains 200 person labels. Each label has one sketch image and two RGB photos. Some examples are shown in Fig. 5. The first and middle rows show RGB photos taken from camera A and camera B. Sketch images are shown on the bottom row. Due to being captured by different cameras, RGB photos have a variety of poses. The sketch images are drawn by five artists. Lots of Volunteers play the role of eyewitnesses to make the sketch images match the real forensic sketch images. They capture the person's characteristics after watching RGB photos for a period. Then the sketch artists draw the sketch images according to the volunteers' descriptions. Totally five painting styles (A $\sim$ E) in this dataset. Details of sketch image labels are in Table I. Same as the settings of the dataset presenter [10], we randomly select 3/4 of all labels for training and 1/4 of all labels for testing from each painting style. There are 150 labels for training and 50 labels for testing. The training set is composed of 150 labels' sketch images and RGB photos. And 50 labels' sketch images and RGB photos are used for the testing set. The settings of training and testing are shown in Table II.

TABLE II

DETAILS OF TRAINING AND TESTING ON SKETCH RE-ID DATASET

Item	Domain	Labels	Sketch image or RGB photo numbers
Training	Sketch	150	150
	RGB	150	300
Testing	Sketch	50	50
	RGB	50	100

TABLE III

DETAILS OF TRAINING/TESTING SET ON CUHK STUDENT DATASET AND CUFSF

Item	CUHK student dataset	CUFSF
Sketch/Photo(Training)	88	944
Sketch/Photo(Testing)	100	250
Total	188	1194

2) Sketch-Photo Face Datasets: We also proved the generalization performance of the proposed model on the CUHK student dataset and CUFSF dataset. The details of training/testing on them are shown in Table III. The CUHK student dataset has 188 labels. Each label has one sketch image and one RGB photo. Sketch images are drawn by the artist and RGB photos are taken in a frontal pose under normal lighting. The training set consists of 88 random labels' sketch images and RGB photos. The testing set is composed of the other 100 labels' sketch images and RGB photos. The CUFSF has 1194 labels. There are also a sketch image and a photo for each label. But photos in the CUFSF are in black-and-white with lighting variations. And sketch images are exaggerated. We employ the MTCNN [44] to detect and align the face. All pictures are cropped throuth the MTCNN [44]. There have been 944 labels' sketch images and photos selected as the training set. The others are used for the testing set.

B. Experiment Setting

1) Implementation Details: The model is built with the Pytorch [45] and trained on double Nvidia Geforce GTX 1080Ti 11 GB GPU. The backbones of two branches are pre-trained on the ImageNet-1k [46] but not on the public Re-id datasets (e.g., Market-1501 [47], Dukemtmc [48] and

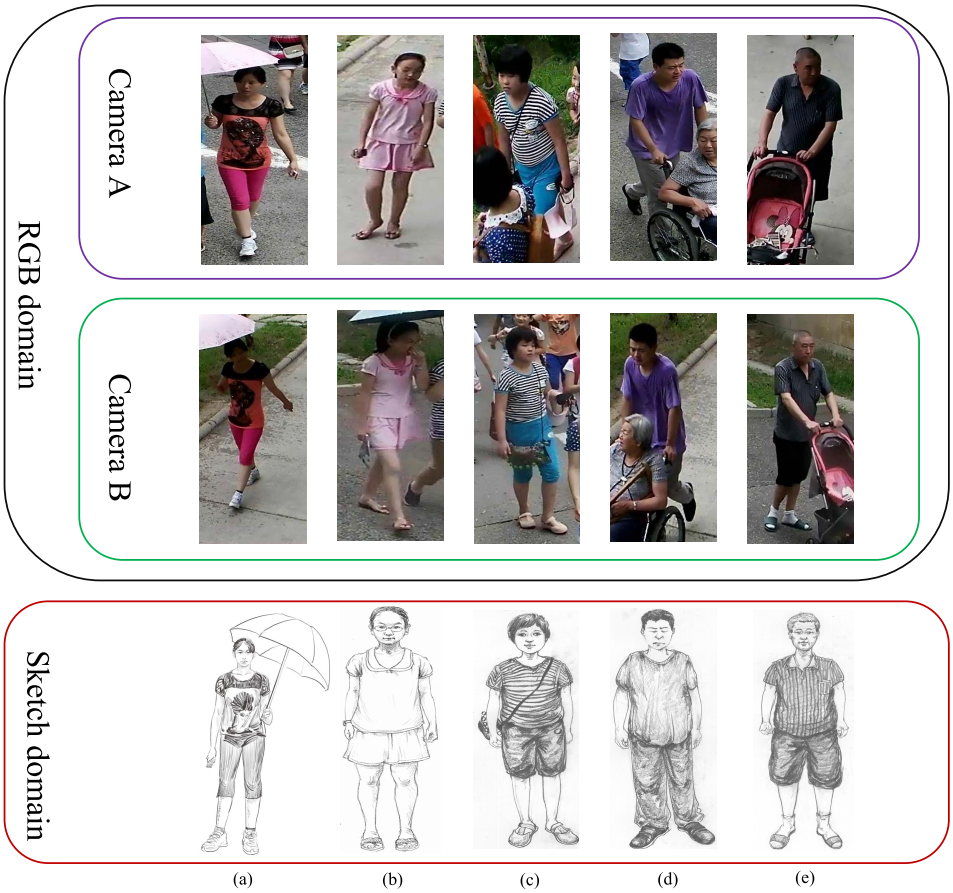

Fig. 5. Some labels' sketch images and RGB photos. The RGB photos are captured by camera A or camear B. (a-e) Human postures are various. (c) RGB photos are affected by complex and messy backgrounds. (d-e) The person is obscured by other objects.

CUHK03 [49]). Then the whole model is fine-tuned to our tasks. We resize all inputs with 224×224 and perform the data augmentation with random horizontal flipping with the probability of 0.5 before training. Different training strategies are adopted on three datasets. On sketch-photo face datasets, we use an Adam optimizer with a learning rate of 6×10^{-5} and train the model for 300 epochs. The batch size of the CUHK student dataset and CUFSF is set to 22 and 30. On the Sketch Re-id dataset, the model is trained with a batch size of 30 for 200 epochs. And we use an Adam optimizer with a learning rate of 1×10^{-4}. The β_1 and β_2 of all Adam optimizer are set to 0.9 and 0.999.

2) Evaluation Metrics: The retrieval accuracy Rank-K is used as the evaluation metric. Rank-K means the percentage of the true match for the given probe image appearing in the first K images. We use Rank-1, Rank-5, Rank-10, and Rank-20 in our experiment to prove the model performance. Because the Sketch Re-id dataset is small, we report the average accuracy value of 10 experiments as the final result to avoid contingency. The train set and test set in each experiment are randomly re-sampled from the Sketch Re-id dataset.

C. Results

1) Comparison With Other Baselines on the Sketch Re-Id Dataset: There are seven baselines used for comparing with our proposed model. The Rank-1, Rank-5, Rank-10, and Rank-20 of each baseline are shown in Table IV. We take the average accuracy of 10 experiments as the result for all baselines. Our model can outperform the previous methods by 3.7% on the Sketch Re-id dataset. Triplet SN [50] is designed for the recognition of the free-hand sketch images. The edge maps extracted from person RGB photos and person sketch images are fed into the Triplet SN in the testing. The worst result indicates that this model cannot extract the domain-invariant features between different domains. GN Siamese [51] model contains GoogleNet [20], which is trained with Siamese and classification loss. GN Siamese can learn the common semantic feature between the sketch

TABLE IV
COMPARISON RESULTS WITH OTHER BASELINES ON THE SKETCH RE-ID DATASET

Model	Rank-1	Rank-5	Rank-10	Rank-20
Triplet SN [50]	9.0%	26.8%	42.2%	65.2%
GN Siamese [51]	28.9%	54.0%	62.4%	78.2%
AFL Net [10]	34.0%	56.3%	72.5%	84.7%
RCD [52]	42.5%	70.0%	87.5%	-
TC-Net [53]	48.1%	73.7%	81.7%	91.5%
MDFL Net [54]	49.0%	70.4%	80.2%	92.0%
UFE [55]	57.1%	79.6%	**89.8%**	93.9%
Our model	**60.8%**	**80.6%**	88.8%	**95.0%**

TABLE V

STANDARD DEVIATION OF THE RANK-K ACCURACY ON THE SKETCH RE-ID DATASET

Item	Rank-1	Rank-5	Rank-10	Rank-20
Avg	60.80%	80.60%	88.80%	95.00%
Standard Deviation	0.0244	0.0241	0.0194	0.0195

TABLE VI

COMPARISON RESULTS WITH OTHER ATTENTION MECHANISMS ON THE SKETCH RE-ID DATASET

Attention mechanism	Rank-1
Non-local [56]	40.75%
SE [57]	56.25%
SA [26]	56.50%
ECA [28]	56.75%
CDA	**60.80%**

domain and RGB domain. AFL [10] is a cross-domain adversarial learning model to jointly learn identity features and domain-invariant features. RCD [52] uses a random homogeneous transformation to realize the modeling of different modal relationships. TC-Net [53] consists of a triplet Siamese network and an auxiliary classification loss to help learn more discriminative features. MDFL [54] learns the invariant features between multiple domains through fusing the multi-level features. UFE [55] designs an unbiased feature extractor and applies the multi-stream classifier, which improves the model's capability of extracting features and bridging the domain gap. Among them, Triplet SN [50], GN Siamese [51] and AFL [10] are pre-trained on the public person Re-ID dataset Market1501 [47]. But RCD [52], TC-Net [53], MDFL [54], UFE [55] and our model can achieve a considerable result but without extra pre-trained on public Re-id dataset. Due to the efficient cross-domain attention mechanism and novel cross-domain center loss, our model can perform better than previous methods. Besides, Table V lists the standard deviation of each Rank-K accuracy, which also proves the validity and stability of our model.

2) Comparison With Other Attention Mechanisms on the Sketch Re-Id Dataset: We use other attention mechanisms to replace our cross-domain attention mechanism and test their performance. SE [57] and ECA [28] are channel attention. SA [26] focuses on spatial information. Non-local [56] implements the self-attention with the convolution. All experiments use our proposed cross-domain center loss. The result is also averaged from 10 experiments. The Rank-1 of each attention mechanism is shown in Table VI. The experiment results show that our cross-domain attention mechanism performs better than the above attention mechanisms in the Sketch Re-id task.

3) Comparison With Other Methods on the Sketch-Photo Face Datasets: To evaluate the generalization ability of our model, we experiment on the CUHK student dataset and CUFSF. The evaluation metrics used are Rank-1. The performance of all state-of-the-art methods and ours are shown in Table VII and Tabel VIII. In the comparison on the CUHK student dataset, LLE [58] is a patch-based synthesis method that can synthesize the target sketch patch from several training sketch patches according to the RGB photos. MRF [11] is a multi-scale model, which can learn the relations among neighboring image patches. DR-GAN [59] performs face frontalization and learns pose-invariant representation. It can take one or multiple images as the input and generate one unified representation. Dual-transfer [60] that uses the dual-transfer FPSS framework is composed of an inter-domain transfer process and an intra-domain transfer process. ODL-CNN [61] is a new IoT-enabled Optimal Deep Learning method. The hyper parameter optimization of ODL-CNN is the Improved Elephant Herd Optimization algorithm. IA CycleGan [62] applies a perceptual loss to supervise the image generation network, which pays more attention to the key facial regions. KD model [63] uses a teacher network to learn the knowledge of the face in the sketch and RGB domain. Then this knowledge is transferred to two student networks designed for the face photo-sketch synthesis task. We also compare our result with some methods on CUFSF. Besides the above methods in [62], [63], and [61], there are other two methods on CUFSF. G-HFR [32] uses the graphical representation and the Markov network to calculate the spatial compatibility between the neighboring image patches. MMTN [40] explore the prototypical style patterns of the reference domain. Our model achieves close performance on the CUFSF dataset and state-of-the-art performance on the CUHK student dataset. All these results prove the validity and the generalization of the cross-domain attention mechanism and center loss.

TABLE VII

COMPARISON RESULTS WITH OTHER BASELINES ON CUHK STUDENT DATASET

Methods	Rank-1
LLE [58]	85.00%
MRF [11]	85.70%
DR-GAN [59]	83.70%
Dual-Transfer [60]	86.30%
ODL-CNN [61]	91.35%
IA CycleGAN [62]	93.62%
KD [63]	93.55%
Our model	**94.00%**

TABLE VIII

COMPARISON RESULTS WITH OTHER BASELINES ON CUFSF

Methods	Rank-1
IA CycleGAN [62]	64.94%
KD [63]	66.37%
ODL-CNN [61]	87.31%
G-HFR [32]	96.00%
MMTN [40]	**97.20%**
Our model	96.80%

TABLE IX

ABLATION RESULTS ON THE SKETCH RE-ID DATASET

CDA		CDC loss	Rank-1
global	local		
×	×	×	49.75%
✓	×	×	52.75%
✓	✓	×	56.00%
×	×	✓	56.25%
✓	✓	✓	**60.80%**

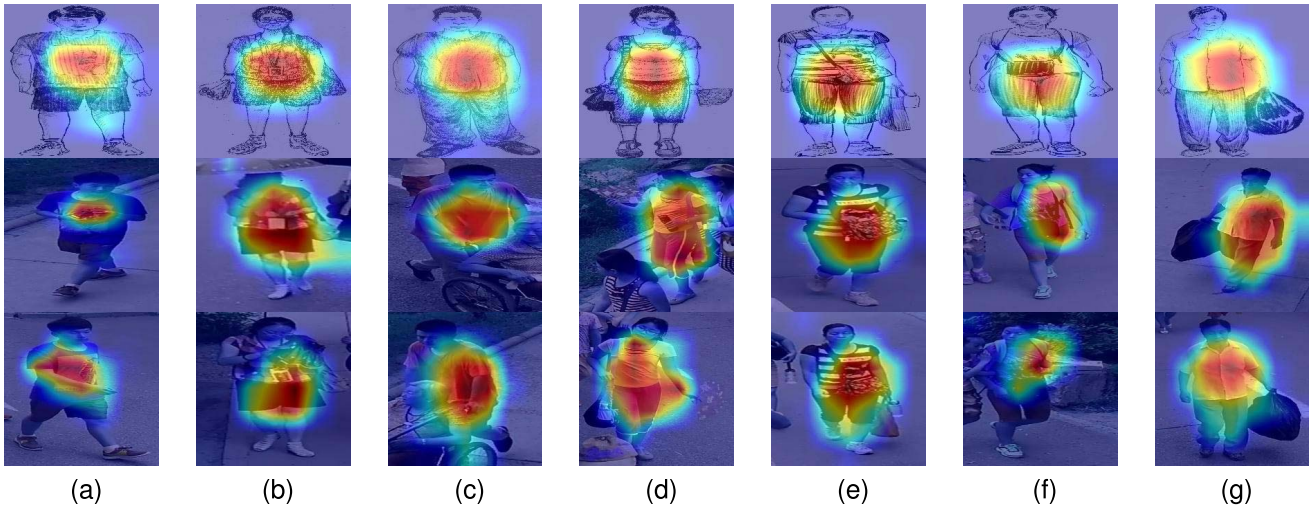

Fig. 6. Grad-CAM visualization of attention maps of sketch images and RGB photos in Sketch Re-id dataset.

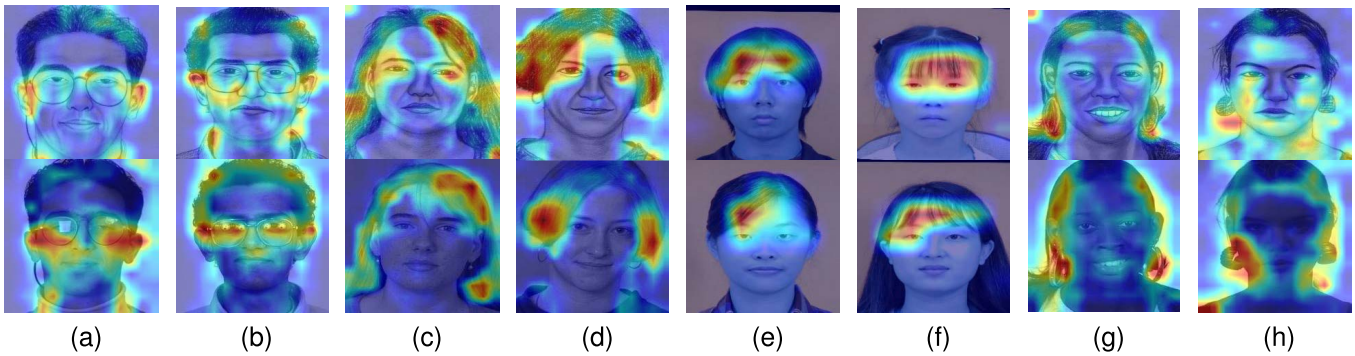

Fig. 7. Grad-CAM visualization of attention maps of sketch images and RGB photos in CUFSF and CUHK student dataset.

4) Ablation Study: We have learned the domain-invariant features through the designed CDA. And the model is optimized by the jointly novel cross-domain center loss and identity features learning losses (classification loss and triplet loss). To prove the validity of the cross-domain attention mechanism and center loss, we perform a series of ablation experiments on the Sketch Re-id dataset. The experiment results are shown in Table IX. The Rank-1 is 49.75% when using the backbone without any other operations. We first add the CDA without the local branch to the backbone and achieve 52.75% accuracy. The complete CDA can make the Rank-1 increase to 56.00%. Then we only use the cross-domain center loss in the training and find that the Rank-1 increases to 56.25%. The above ablation results confirm that our novel cross-domain attention mechanism and center loss can make the model perform better on the Sketch Re-id task.

D. Visualization and Qualitative Analysis

Through Grad-CAM [64], we visualize the sketch image feature maps and RGB photo feature maps after the CDA in our model. Both Sketch Re-id dataset and sketch-photo face datasets generate 7 heatmaps. As shown in Fig. 6 and Fig. 7, our model extracts greater local features in the RGB photo and reduces the difference between the sketch images and RGB photos effectively. Besides the visualization results, we list the retrieval results on three datasets in Fig. 8 and Fig. 9. The results in the Sketch Re-id dataset also be compared to the model without the CDA. The correct retrieval results and incorrect retrieval results are highlighted with a dotted green line and a red dotted line respectively.

1) Sketch Re-Id Dataset: The heatmaps of the Re-id dataset are shown in Fig. 6, we can notice that our model can reduce interference in the background and focus on similar areas in the sketch images and RGB photos. In columns (a) and (b), the model focuses more on the logo and badges, which are on the garment on the chest. From columns (c) and (d), the RGB photos contain two persons. Yet based on the guide of the sketch images, the model can focus on the right person through the same clothes features in the sketch image. Besides, the model can also track the local feature of the pedestrians. In columns (e) and (f), due to the pedestrians' posture being different in the RGB photos and sketch images, the backpack position of the pedestrians is different. Our model can find the backpack in the RGB photos according to the backpack feature in the sketch images. But sometimes the position of objects held by pedestrians may change a lot. Taking the sketch image in column (g) for example, the garbage bag is in the person's left hand, but it appears on the person's left or right hand in the real RGB photo. The local feature like this cannot be an important feature for recognition. So, the model takes the

Fig. 8. Retrieval results on the Sketch Re-id dataset. (a) the complete model (b) the model without the CDA.

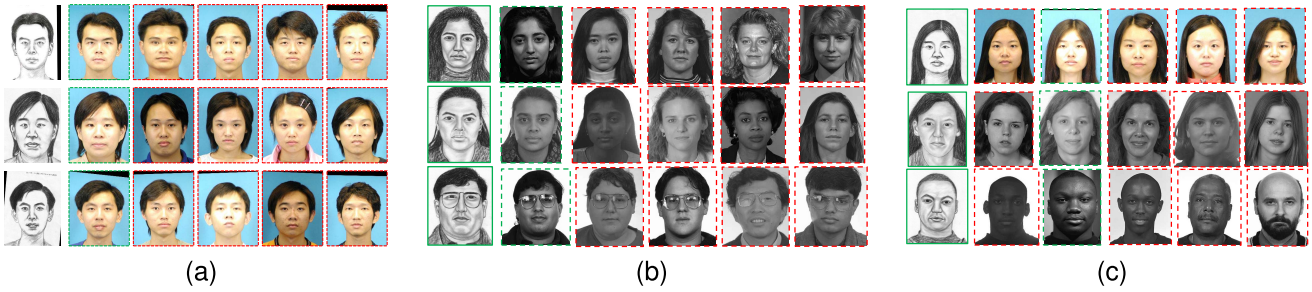

Fig. 9. Retrieval results of our model in the CUHK student dataset and CUFSF.

clothes and arms of pedestrians as distinguishing features. The retrieval results are shown in Fig. 8. We compare the complete model with or without the CDA. As shown in Fig. 8 (a), the correct person in the gallery can be found in the Rank-1 with a complete model. But the result in Fig. 8 (b) shows that the model without the CDA can only find the correct person in Rank-3 or Rank-4.

2) Sketch-Photo Face Datasets: The heatmaps of the sketch-photo face datasets are shown in Fig. 7. The (a), (b), (c), (d), (g), and (h) are from CUFSF, and the (e) and (f) are from the CUHK student dataset. Our model has significant performance on these datasets. The designed CDA makes the backbone focus on the local details of the human face. In (a) and (b), the model captures the differences between glasses and through them distinguishes different people. And in (c) to (f), the model focuses on the hairstyles and the hair on the forehead. The girl's hairstyle is completely different in (c) and (d). There is also a slight difference between the person's hair on the forehead in (e) and (f). As for (g) and (h), the model focuses on the earrings. We can find that the earrings in (g) are a little big than those in (h). It is the cross-domain attention mechanism that makes the model capture the above features. The Fig. 9 shows some retrieval results. The (a) and (b) show the correct Rank-1 retrieval results on the CUHK student dataset and CUFSF. It can be found that our model can find the correct retrieval results in Rank-1 whether the photo in the gallery is RGB or black-and-white. We also list some failure cases in (c). In the first and third row, the light on each face and the background color is different. These differences cause some difficulties for the model. In addition, some similar features also cause interference when the model focuses on the same local regions, like the nose and mouth in the second row. Although our model does not find the correct target on Rank-1, it can still be found on Rank-2.

V. Conclusion

In the Sketch Re-id task, due to the query being sketch domain and the gallery being RGB domain, it is difficult for the model to extract the identified domain-invariant features. Instead of previous methods extracting the features of the two domains respectively for optimization, we propose a novel idea that uses the features of the sketch domain to guide the extraction of the RGB features. Based on this idea, we design the cross-domain attention (CDA) mechanism. Specifically, the CDA has two different branches: the global branch and the local branch. The difference between the two branches is how to split the feature maps and get the final attention weight. Besides, we use the cross-domain center loss (CDC) to catch the domain-invariant feature. Through the change of center point in common space, the difference between the

sketch domain and RGB domain can be effectively reduced. In addition to the Sketch Re-id task, we perform experiments in the cross-domain face recognition task to prove the generalization of our model. Experimental results show that our model achieves the state-of-the-art performance on the Sketch Re-id dataset and two sketch-photo face recognition datasets. In general, our model is universal for sketch-based cross-domain pedestrian Re-id tasks and face recognition tasks.

REFERENCES

[1] L. Zheng, H. Zhang, S. Sun, M. Chandraker, Y. Yang, and Q. Tian, "Person re-identification in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1367–1376.

[2] S. Tang, M. Andriluka, B. Andres, and B. Schiele, "Multiple people tracking by lifted multicut and person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3539–3548.

[3] H. Kiani Galoogahi and T. Sim, "Face photo retrieval by sketch example," in *Proc. 20th ACM Int. Conf. Multimedia*, 2012, pp. 949–952.

[4] W. Zhang, X. Wang, and X. Tang, "Coupled information-theoretic encoding for face photo-sketch recognition," in *Proc. CVPR*, Jun. 2011, pp. 513–520.

[5] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[6] V. S. Vibashan, J. M. J. Valanarasu, P. Oza, and V. M. Patel, "Image fusion transformer," 2021, *arXiv:2107.09011*.

[7] W. Kim, B. Son, and I. Kim, "ViLT: Vision- and-language transformer without convolution or region supervision," 2021, *arXiv:2102.03334*.

[8] R. Bose, S. Pande, and B. Banerjee, "Two headed dragons: Multimodal fusion and cross modal transactions," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2021, pp. 2893–2897.

[9] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014, *arXiv:1409.0473*.

[10] L. Pang, Y. Wang, Y.-Z. Song, T. Huang, and Y. Tian, "Cross-domain adversarial feature learning for sketch re-identification," in *Proc. 26th ACM Int. Conf. Multimedia*, Oct. 2018, pp. 609–617, doi: 10.1145/3240508.3240606.

[11] X. Wang and X. Tang, "Face photo-sketch synthesis and recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 11, pp. 1955–1967, Nov. 2009.

[12] X. Zhang *et al.*, "AlignedReID: Surpassing human-level performance in person re-identification," 2017, *arXiv:1711.08184*.

[13] Y.-C. Chen, X. Zhu, W.-S. Zheng, and J.-H. Lai, "Person re-identification by camera correlation aware feature augmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 2, pp. 392–408, Feb. 2018.

[14] L. Zheng, Y. Yang, and A. G. Hauptmann, "Person re-identification: Past, present and future," 2016, *arXiv:1610.02984*.

[15] T. Matsukawa and E. Suzuki, "Person re-identification using CNN features learned from combination of attributes," in *Proc. 23rd Int. Conf. Pattern Recognit. (ICPR)*, Dec. 2016, pp. 2428–2433.

[16] Y. Lin *et al.*, "Improving person re-identification by attribute and identity learning," *Pattern Recognit.*, vol. 95, pp. 151–161, Nov. 2019.

[17] H. Liu, J. Feng, M. Qi, J. Jiang, and S. Yan, "End-to-end comparative attention networks for person re-identification," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3492–3506, May 2017.

[18] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng, "Person re-identification by multi-channel parts-based CNN with improved triplet loss function," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1335–1344.

[19] R. R. Varior, M. Haloi, and G. Wang, "Gated Siamese convolutional neural network architecture for human re-identification," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 791–808.

[20] H. Yao, S. Zhang, R. Hong, Y. Zhang, C. Xu, and Q. Tian, "Deep representation learning with part loss for person re-identification," *IEEE Trans. Image Process.*, vol. 28, no. 6, pp. 2860–2871, Jun. 2019.

[21] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, "Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018 pp. 480–496.

[22] A. Dosovitskiy *et al.*, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[23] Z. Liu *et al.*, "Swin transformer: Hierarchical vision transformer using shifted windows," 2021, *arXiv:2103.14030*.

[24] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 213–229.

[25] J. Wang *et al.*, "M2TR: Multi-modal multi-scale transformers for deepfake detection," 2021, *arXiv:2104.09770*.

[26] H. Wang, Y. Fan, Z. Wang, L. Jiao, and B. Schiele, "Parameter-free spatial attention network for person re-identification," 2018, *arXiv:1811.12150*.

[27] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2011–2023, Aug. 2020.

[28] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: Efficient channel attention for deep convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11531–11539.

[29] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 3–19.

[30] X. Ma *et al.*, "Learning connected attentions for convolutional neural networks," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2021, pp. 1–6.

[31] J. Fu *et al.*, "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3141–3149.

[32] C. Peng, X. Gao, N. Wang, and J. Li, "Graphical representation for heterogeneous face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 2, pp. 301–312, Feb. 2017.

[33] D. Lin and X. Tang, "Inter-modality face recognition," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2006, pp. 13–26.

[34] J. Huo, Y. Gao, Y. Shi, and H. Yin, "Cross-modal metric learning for AUC optimization," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 10, pp. 4844–4856, Oct. 2018.

[35] M. Kan, S. Shan, H. Zhang, S. Lao, and X. Chen, "Multi-view discriminant analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 1, pp. 188–194, Jan. 2016.

[36] C. Peng, N. Wang, J. Li, and X. Gao, "Soft semantic representation for cross-domain face recognition," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 346–360, 2020.

[37] Y. Jin, J. Lu, and Q. Ruan, "Coupled discriminative feature learning for heterogeneous face recognition," *IEEE Trans. Inf. Forensics Security*, vol. 10, no. 3, pp. 640–652, Mar. 2015.

[38] X. Tang and X. Wang, "Face sketch synthesis and recognition," in *Proc. 9th IEEE Int. Conf. Comput. Vis.*, Oct. 2003, pp. 687–694.

[39] M. Zhang, Y. Li, N. Wang, Y. Chi, and X. Gao, "Cascaded face sketch synthesis under various illuminations," *IEEE Trans. Image Process.*, vol. 29, pp. 1507–1521, 2019.

[40] M. Luo, H. Wu, H. Huang, W. He, and R. He, "Memory-modulated transformer network for heterogeneous face recognition," *IEEE Trans. Inf. Forensics Security*, vol. 17, pp. 2095–2109, 2022.

[41] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 499–515.

[42] L. Jing, E. Vahdani, J. Tan, and Y. Tian, "Cross-modal center loss for 3D cross-modal retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 3142–3151.

[43] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," 2017, *arXiv:1703.07737*.

[44] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016.

[45] A. Paszke *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, Dec. 2019, pp. 8026–8037.

[46] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[47] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1116–1124.

[48] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 17–35.

[49] W. Li, R. Zhao, T. Xiao, and X. Wang, "DeepReID: Deep filter pairing neural network for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 152–159.

[50] Q. Yu, F. Liu, Y.-Z. Song, T. Xiang, T. M. Hospedales, and C. C. Loy, "Sketch me that shoe," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 799–807.
[51] P. Sangkloy, N. Burnell, C. Ham, and J. Hays, "The sketchy database: Learning to retrieve badly drawn bunnies," *ACM Trans. Graph.*, vol. 35, no. 4, pp. 1–12, Jul. 2016.
[52] Y. Gong, L. Huang, and L. Chen, "Eliminate deviation with deviation for data augmentation and a general multi-modal data learning method," 2021, *arXiv:2101.08533*.
[53] H. Lin, Y. Fu, P. Lu, S. Gong, X. Xue, and Y.-G. Jiang, "TC-Net for iSBIR: Triplet classification network for instance-level sketch based image retrieval," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 1676–1684.
[54] S. Gui, Y. Zhu, X. Qin, and X. Ling, "Learning multi-level domain invariant features for sketch re-identification," *Neurocomputing*, vol. 403, pp. 294–303, Aug. 2020.
[55] B. Yuan, B. Chen, Z. Tan, X. Shao, and B.-K. Bao, "Unbiased feature enhancement framework for cross-modality person re-identification," *Multimedia Syst.*, vol. 28, no. 3, pp. 749–759, 2022.
[56] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7794–7803.
[57] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.
[58] Q. Liu, X. Tang, H. Jin, H. Lu, and S. Ma, "A nonlinear approach for face sketch synthesis and recognition," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Jun. 2005, pp. 1005–1010.
[59] L. Tran, X. Yin, and X. Liu, "Disentangled representation learning GAN for pose-invariant face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1415–1424.
[60] M. Zhang, R. Wang, J. Gao, X. Li, and D. Tao, "Dual-transfer face sketch–photo synthesis," *IEEE Trans. Image Process.*, vol. 28, no. 2, pp. 642–657, Sep. 2019.
[61] M. Elhoseny, M. M. Selim, and K. Shankar, "Optimal deep learning based convolution neural network for digital forensics face sketch synthesis in Internet of Things (IoT)," *Int. J. Mach. Learn. Cybern.*, vol. 12, pp. 3249–3260, Jul. 2020.
[62] Y. Fang, W. Deng, J. Du, and J. Hu, "Identity-aware cycleGAN for face photo-sketch synthesis and recognition," *Pattern Recognit.*, vol. 102, Jun. 2020, Art. no. 107249.
[63] M. Zhu, J. Li, N. Wang, and X. Gao, "Knowledge distillation for face photo–sketch synthesis," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 2, pp. 893–906, Feb. 2022.
[64] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.

Fengyao Zhu received the B.S. degree from Nanjing Agricultural University in 2020. He is currently pursuing the master's degree with the School of Information Science and Engineering, East China University of Science and Technology. His research interests include deep learning, image processing, and pattern recognition.

Yu Zhu (Member, IEEE) received the Ph.D. degree in optical engineering from the Nanjing University of Science and Technology, China, in 1999. She is currently a Professor with the Department of Electronics and Communication Engineering, East China University of Science and Technology. She has published more than 100 papers in journals and conferences. Her research interests include image processing, computer vision, multimedia communication, and deep learning.

Xiaoben Jiang is currently pursuing the Ph.D. degree with the East China University of Science and Technology. His experience includes the denoising method on chest X-ray images and CT images and the detection of COVID-19 cases from denoised CXR images. He has published in journals in the crossing field of medical science and computer vision. He has been involved in publicly and privately funded projects. His current research interests include digital image processing and computer vision.

Jiongyao Ye received the M.B. and Ph.D. degrees from Waseda University, Japan, in 2005 and 2011, respectively. From 2005 to 2008, he worked as a Senior Researcher with SONY Inc. He is currently an Associate Professor with the East China University of Science and Technology. His research interests include IC design, image signal process, CMOS image sensor, and low power design. He twice won IEEE best conference paper awards and four IC layout patented.