Cross-Compatible Embedding and Semantic Consistent Feature Construction for Sketch Re-identification

Yafei Zhang
zyfeimail@163.com
Faculty of Information Engineering and Automation,
Kunming University of Science and Technology
Kunming, China

Yongzeng Wang
wangyongzeng@stu.kust.edu.cn
Faculty of Information Engineering and Automation,
Kunming University of Science and Technology
Kunming, China

Huafeng Li*
hfchina99@163.com
Faculty of Information Engineering and Automation,
Kunming University of Science and Technology
Kunming, China

Shuang Li†
shuangli936@gmail.com
Faculty of Information Engineering and Automation,
Kunming University of Science and Technology
Kunming, China

ABSTRACT

Sketch re-identification (Re-ID) refers to using sketches of pedestrians to retrieve their corresponding photos from surveillance videos. It can track pedestrians according to the sketches drawn based on eyewitnesses without querying pedestrian photos. Although the Sketch Re-ID concept has been proposed, the gap between the sketch and the photo still greatly hinders pedestrian identity matching. Based on the idea of transplantation without rejection, we propose a Cross-Compatible Embedding (CCE) approach to narrow the gap. A Semantic Consistent Feature Construction (SCFC) scheme is simultaneously presented to enhance feature discrimination. Under the guidance of identity consistency, the CCE performs cross modal interchange at the local token level in the Transformer framework, enabling the model to extract modal-compatible features. The SCFC improves the representation ability of features by handling the inconsistency of information in the same location of the sketch and the corresponding pedestrian photo. The SCFC scheme divides the local tokens of pedestrian images with different modes into different groups and assigns specific semantic information to each group for constructing a semantic consistent global feature representation. Experiments on the public Sketch Re-ID dataset confirm the effectiveness of the proposed method and its superiority over existing methods. Experiments on Sketch-based image retrieval datasets QMUL-Shoe-v2 and QMUL-Chair-v2 are conducted to assess the method's generalization. The results show that the proposed method outperforms the state-of-the-art works compared. The source code of the proposed method is available at: https://github.com/lhf12278/CCSC.

*Corresponding author.
†Corresponding author.

MM '22, October 10–14, 2022, Lisboa, Portugal

ACM ISBN 978-1-4503-9203-7/22/10…$15.00
https://doi.org/10.1145/3503161.3548224

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**; • **Information systems** → *Information retrieval*.

KEYWORDS

Sketch Re-identification, Cross-Compatible Embedding, Cross transplantation, Semantic Consistent Features.

ACM Reference Format:
Yafei Zhang, Yongzeng Wang, Huafeng Li, and Shuang Li. 2022. Cross-Compatible Embedding and Semantic Consistent Feature Construction for Sketch Re-identification. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22), October 10–14, 2022, Lisboa, Portugal.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3503161.3548224

1 INTRODUCTION

Person re-identification (Re-ID) is commonly used for judging whether two pedestrians are the same identity in non-overlapping surveillance cameras. Person Re-ID has attracted more and more attention since it provides valuable technical support for tracking suspects and searching for missing persons. However, existing person Re-ID models are mainly concerned with matching pedestrian photos. Practically, all areas cannot be entirely covered by monitoring equipment, which makes only eyewitnesses but no images of suspect more likely occur. To this end, the Sketch Re-ID is proposed for matching the sketch drawn by a professional artist according to the description of an eyewitness to pedestrian photos taken by surveillance video.

However, due to the impact of modal gap between sketches and photos, existing Re-ID models designed for photo-to-photo matching cannot be directly used to solve the matching between sketch and person photos. As shown in Figure 1, a sketch merely contains person's outline information while a person photo not only shows clear outlines, but also rich color and complex background information. Besides, a sketch usually shows the person with the frontal view while a person photo could come from any camera and be affected by multiple factors such as camera view, pedestrian pose and occlusion. In addition, owing to inaccurate pedestrian detection [8, 12], the information at the same spatial location does not correspond between photos and sketches (as can be seen in Figure 1, the information in the boxes at the same location are

Figure 1: Sketches of pedestrians and their corresponding photos. From these image pairs, we can see that the information is inconsistent at the same location.

not match). Thus, Sketch Re-ID could be more challenging than traditional person Re-ID.

In order to solve the modal discrepancy problem, Pang et al. [18] proposed the Sketch Re-ID and raised a cross-domain adversarial learning method to acquire the modal-invariant features. However, the adversarial learning would easily lead to non-common discriminative features loss, which affects the model's generalization ability. Gui et al. [9] jointly mapped photos and sketches into a common embedded space to narrow the modal gap between them. Although the performance of this method was improved compared with that of [18], the modal discrepancy remains a thorny issue. Moreover, the above methods do not consider non-corresponding information at the same spatial location between sketch and photo for pedestrian identity matching, which limits further performance improvement. Meanwhile, for the Re-ID tasks, non-tailored feature extraction networks tend to focus on the most discriminative region of certain parts, causing a direct effect on the feature discrimination if this area is occluded [16].

In this paper, we develop a new cross-compatibility and semantic consistent feature construction Transformer framework to address the above problems. It contains a cross-compatible module and a semantic consistency feature construction module. The former is mainly used to narrow the gap between modalities. The latter is for mitigating the impact on feature discrimination caused by the non-corresponding information at the same spatial location and promoting the feature extraction network to extract discriminative features from different parts of the pedestrian. Therefore, the method in this paper can improve the discrimination of features while alleviating the modal discrepancy.

The idea of the cross-compatible module is transplantation without rejection, which draws on the experience of organ transplantation. In the process, if the classifier trained with the specific modal features does not repel after the features of different modalities are cross transplanted (i.e., the features obtained after transplantation can still be correctly classified by the original classifier), it can be deemed that the corresponding transplanted features are compatible with the current modal features. The compatibility indicates that the gap between different modalities has been narrowed. For semantic consistent feature construction, the local tokens of different modalities are grouped and sent into the Transformer layer to obtain the class tokens of each group. In this way, each group of one image can be assigned specific semantic information, and different groups can be assigned different semantic information. Meanwhile, the corresponding class tokens between different modalities hold the same semantic information. With those class tokens given semantic information, the semantic consistent feature representation can be achieved by combining these semantic class tokens into new classes in order. Moreover, there are class differences between different groups, which encourages the network to extract features from different regions of the pedestrian image. It helps to improve the feature representation. The contributions of this work are summarized as follows.

(1) Inspired by the idea of transplantation without rejection, a cross-compatible learning mechanism is built to narrow the gap between different modal features. Compared with the existing modal gap reduction methods, the proposed method can mitigate the impact of modal differences and avoid the loss of discriminative features.
(2) A semantic consistent feature construction mechanism is designed in the Transformer framework to deal with the non-corresponding information at the same spatial location of different modalities and the problem that the feature extraction network tends to pay attention to a particular local area of pedestrian image. The feature semantic consistent is achieved by assigning different semantic information to intra-modality class tokens of different groups while assigning the same semantic category to the corresponding inter-modality class tokens.
(3) The results on the public Sketch Re-ID dataset show that the proposed method achieves better performance than the state-of-the-art methods compared. It also shows satisfying generalizability on Sketch-based image retrieval datasets.

2 RELATED WORK

2.1 Sketch-based Image Retrieval

Sketch Re-ID uses sketches of pedestrians to retrieve their corresponding photos from the surveillance videos. In 2018, Lu et al. [18] proposed the Sketch Re-ID task for the first time and contributed the first Sketch Re-ID dataset. In this method, they used adversarial learning to narrow the gap between sketches and pedestrian photos. However, the dataset is small because those sketches require much human labor. Based on the dataset, Gui et al. [9] used a triplet classification network as the backbone to reduce the gap between sketches and photos by embedding gradient flip layers. Although the method is effective, the challenge of matching sketches with photos is still not effectively solved. Due to the limited size of the sketch-pedestrian dataset, the related works on Sketch Re-ID are scarce.

On the other hand, sketch image retrieval attracts more attention because the technology does not require high-quality sketches, making it easier to build datasets for model performance validation. In particular, Song et al. [22] proposed a fine-grained sketch-based

image retrieval model. It embedded sketches and images into the same space to narrow the gap. Furthermore, Song et al. [21] employed spatial semantic attention modeling to solve the problem of bi-modal feature vectors misalignment among elements through higher-order energy function. Lin et al. [15] developed a triple classification network composed of a triplet Siamese network and classification loss to match sketches with photos. Zhang et al. [6] used the idea of Generative Adversarial Networks (GAN) to narrow the modal gap by making the discriminator unable to distinguish the modal information of the feature. Bhunia et al. [2] constructed a new fine-grained sketch-based image retrieval framework to retrieve the target photo with a sketch with as few strokes as possible to reduce the time consumption of sketch depiction. At the same time, a dynamic design is carried out to start image retrieval as soon as the user starts drawing.

Although the above methods are effective, they may still fail to alleviate the modal gap very well, because the elimination of modal gaps is still an open problem. At the same time, the existing methods do not fully consider the feature representation capability, limiting the further improvement of model performance. Besides, despite the fact that the Sketch Re-ID belongs to image retrieval, it faces more significant challenges due to its higher requirements in accuracy than other sketch image retrieval. This paper further solves the above problems by constructing a transplantation without rejection mechanism and a semantic consistent feature construction scheme.

2.2 Transformer in Person Re-ID

The Transformer was proposed by Vaswani et al. [25] for machine translation. Unlike the convolutional neural networks, it extracts features from the entire input through a self-attention mechanism. Inspired by the Transformer's success in natural language processing, the Transformer is widely applied to the field of computer vision such as object detection [3], semantic segmentation [27], image classification [4], image processing [5] and object tracking [24]. In person Re-ID, He et al. [10] proposed a framework based on the Transformer for the first time. Ma et al. [17] developed an inter- and intra-part relational Transformer under posture guidance to identify obstructed pedestrians, where the Transformer is used to establish part-aware long-term correlations. Li et al. [13] built an end-to-end Part-Aware Transformer to alleviate the impact of occlusion on pedestrian identity matching. Liao et al. [14] introduced the query-gallery concatenation to VIT and query-gallery cross-attention to the vanilla Transformer, respectively, for their lack of image-to-image attention. Zhu et al. [33] raised an automatically aligned Transformer, which brought the same partitioned pedestrian parts together by introducing clustering to achieve semantic feature alignment within the Transformer framework. Unlike the existing methods, this work implements interactive transplantation among different Transformer layers to learn compatible features, narrowing the modal gap between sketches and photos.

3 METHODOLOGY

The overall structure of the proposed method contains two parts: 1) cross-compatible feature learning (CCFL), and 2) the semantic consistent feature construction (SCFC), as illustrated in Figure 2, where the CCFL also includes pedestrian photo feature extraction, sketch feature extraction and cross transplantation (CT) of different modality features. The feature extraction network is composed of 12 Transformer layers. The CCFL narrows the modal gap between sketches and photos. The SCFC constructs a semantic consistent global class token by combining local class tokens with the different semantic information for pedestrian identity matching.

3.1 Feature Extraction

Given a gallery-photo set of n person photos $\tilde{\boldsymbol{X}}^p = \{\tilde{\boldsymbol{x}}_1^p, \tilde{\boldsymbol{x}}_2^p, \cdots, \tilde{\boldsymbol{x}}_n^p\}$ and a probe-sketch set of m person sketches $\tilde{\boldsymbol{X}}^s = \{\tilde{\boldsymbol{x}}_1^s, \tilde{\boldsymbol{x}}_2^s, \cdots, \tilde{\boldsymbol{x}}_m^s\}$. Since the Transformer can only process sequential data, the images need to be partitioned. We first use a $P \times P$ sliding window with stride S to divide each source image into N equal size patches, and then convert each patch into a D-dimensional vector by the linear projection of flattened patches [10]. The i-th image $\tilde{\boldsymbol{x}}_i^p$ and $\tilde{\boldsymbol{x}}_i^s$ are represented by $\boldsymbol{x}_{i,1}^p, \boldsymbol{x}_{i,2}^p, \cdots, \boldsymbol{x}_{i,N}^p$ and $\boldsymbol{x}_{i,1}^s, \boldsymbol{x}_{i,2}^s, \cdots, \boldsymbol{x}_{i,N}^s$, respectively. After adding position $\boldsymbol{P}^p$ and $\boldsymbol{P}^s$, we have:

$$\begin{aligned} \boldsymbol{F}_{i,0}^p &= [\boldsymbol{x}_{i,cls}^{p,0}; \boldsymbol{x}_{i,1}^p; \boldsymbol{x}_{i,2}^p; \cdots; \boldsymbol{x}_{i,N}^p] + \boldsymbol{P}^p \\ &= [\boldsymbol{f}_{i,cls}^{p,0}; \boldsymbol{f}_{i,1}^{p,0}; \boldsymbol{f}_{i,2}^{p,0}; \cdots; \boldsymbol{f}_{i,N}^{p,0}] \end{aligned} \tag{1}$$

$$\begin{aligned} \boldsymbol{F}_{i,0}^s &= [\boldsymbol{x}_{i,cls}^{s,0}; \boldsymbol{x}_{i,1}^s; \boldsymbol{x}_{i,2}^s; \cdots; \boldsymbol{x}_{i,N}^s] + \boldsymbol{P}^s \\ &= [\boldsymbol{f}_{i,cls}^{s,0}; \boldsymbol{f}_{i,1}^{s,0}; \boldsymbol{f}_{i,2}^{s,0}; \cdots; \boldsymbol{f}_{i,N}^{s,0}] \end{aligned} \tag{2}$$

where $\boldsymbol{P}^p \in \mathbb{R}^{(N+1)\times D}$, $\boldsymbol{P}^s \in \mathbb{R}^{(N+1)\times D}$. $\boldsymbol{x}_{i,cls}^{p,0}$ and $\boldsymbol{x}_{i,cls}^{s,0}$ are the class tokens with each element equals to 0 of $\tilde{\boldsymbol{x}}_i^p$ and $\tilde{\boldsymbol{x}}_i^s$ before adding the position embedding. $\boldsymbol{f}_{i,cls}^{p,0}$ and $\boldsymbol{f}_{i,cls}^{s,0}$ are the class tokens of $\tilde{\boldsymbol{x}}_i^p$ and $\tilde{\boldsymbol{x}}_i^s$ after the position information is added, respectively.

Suppose that the network for extracting pedestrian photo features is $\boldsymbol{E}_p$, and the network for extracting sketch features is $\boldsymbol{E}_s$. Both of them have $L(L = 12)$ Transformer layers. After $\boldsymbol{F}_{i,0}^p$ and $\boldsymbol{F}_{i,0}^s$ are fed into $\boldsymbol{E}_p$ and $\boldsymbol{E}_s$ respectively, the corresponding outputs of $\boldsymbol{E}_p$ and $\boldsymbol{E}_s$ are denoted as $\boldsymbol{f}_{i,cls}^p$ and $\boldsymbol{f}_{i,cls}^s$. To ensure that $\boldsymbol{f}_{i,cls}^p$ and $\boldsymbol{f}_{i,cls}^s$ can have strong discrimination, we use the following cross-entropy loss:

$$\mathcal{L}_{id}^p(\boldsymbol{E}_p, \boldsymbol{W}_p) = \sum_{i=1}^{B} -\boldsymbol{y}_i^p log(\boldsymbol{W}_p(\boldsymbol{f}_{i,cls}^p)) \tag{3}$$

$$\mathcal{L}_{id}^s(\boldsymbol{E}_s, \boldsymbol{W}_s) = \sum_{i=1}^{B} -\boldsymbol{y}_i^s log(\boldsymbol{W}_s(\boldsymbol{f}_{i,cls}^s)) \tag{4}$$

and the triplet loss to optimize the network parameters:

$$\mathcal{L}_{tri}^p(\boldsymbol{E}_p) = \sum_{i=1}^{B} [\|\boldsymbol{f}_{i,cls}^p - \boldsymbol{f}_{p,i,cls}^s\|_2^2 - \|\boldsymbol{f}_{i,cls}^p - \boldsymbol{f}_{n,i,cls}^s\|_2^2 + m]_+ \tag{5}$$

$$\mathcal{L}_{tri}^s(\boldsymbol{E}_s) = \sum_{i=1}^{B} [\|\boldsymbol{f}_{i,cls}^s - \boldsymbol{f}_{p,i,cls}^p\|_2^2 - \|\boldsymbol{f}_{i,cls}^s - \boldsymbol{f}_{n,i,cls}^p\|_2^2 + m]_+ \tag{6}$$

where $\boldsymbol{W}_p$ and $\boldsymbol{W}_s$ are the pedestrian identity classifiers corresponding to pedestrian photos and sketches. $\boldsymbol{f}_{i,cls}^p$ and $\boldsymbol{f}_{i,cls}^s$ are the class tokens of pedestrian photo and sketch output by the last Transformer layer, respectively. $\boldsymbol{y}_i^p$ and $\boldsymbol{y}_i^s$ are the identity labels of $\boldsymbol{f}_{i,cls}^p$ and $\boldsymbol{f}_{i,cls}^s$. $\boldsymbol{f}_{p,i,cls}^p(\boldsymbol{f}_{p,i,cls}^s)$ and $\boldsymbol{f}_{n,i,cls}^s(\boldsymbol{f}_{n,i,cls}^p)$ are corresponding class tokens of cross-modal hard positive sample and hard negative sample, respectively. m is a constant. B is the batch size. $[x]_+ = \max(0, x)$.

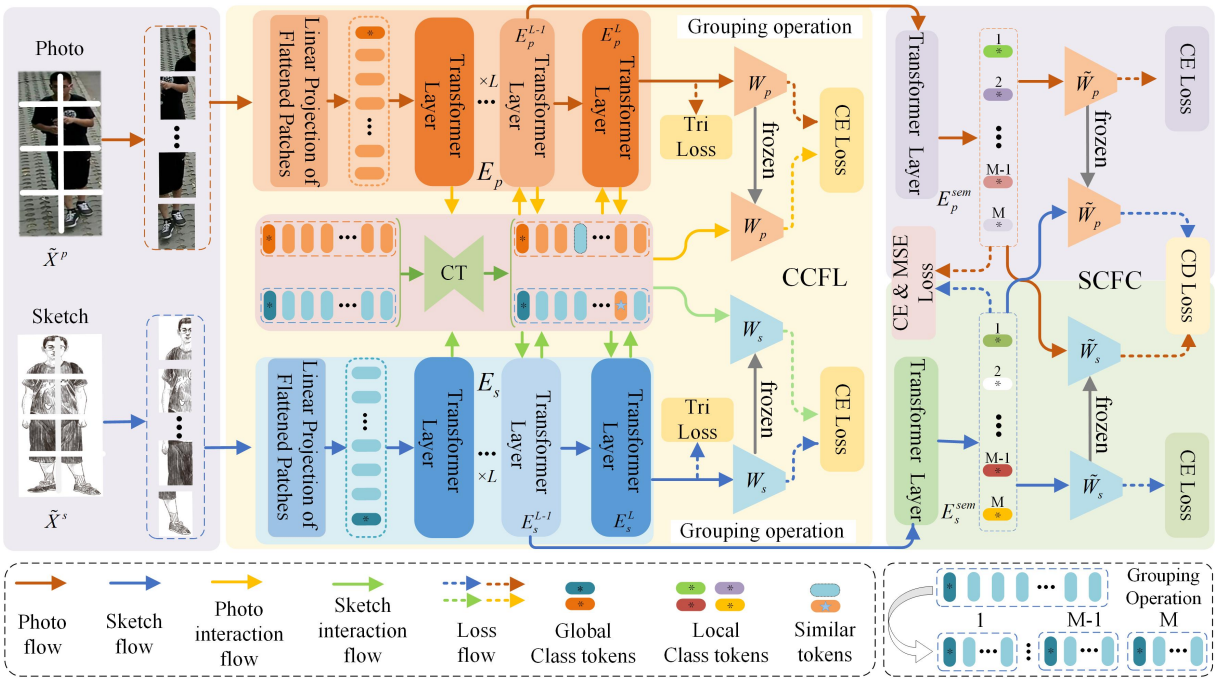

Figure 2: Overall framework of the proposed method. In the training stage, the source images are first used to train feature extraction network and pedestrian identity classifier of each modal features, and then the feature extraction network continues to be trained with the participation of CCFL and SCFC to improve its robustness to modal discrepancy. Particularly, the CCFL obtains new features through the cross transplantation (CT) of different modality features, and feeds them into the trained identity classifier of a specific modality, so that the classifier can classify those features correctly. The SCFC groups local tokens and assigns them different semantic categories according to their groups. Then, the class tokens are used to construct semantic consistent global features.

3.2 Cross-Compatible Feature Learning

The modal gap between sketches and photos is one of the key factors affecting pedestrian identity matching. To solve this problem, we propose a CCFL mechanism based on the idea of transplantation without rejection, which can promote the network to learn modal-invariant features. The proposed cross transplantation structure is shown in Figure 3.

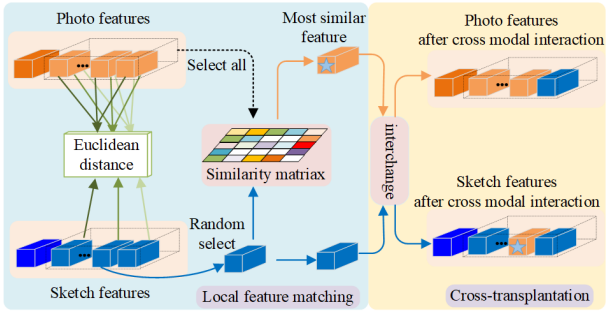

Figure 3: Proposed cross transplantation structure.

Let $F^p_{i,l}$ and $F^s_{i,l}$ be the outputs of the l-th Transformer layer, and they can be formulated as:

$$F^p_{i,l} = E^l_p(F^p_{i,0}) = [f^{p,l}_{i,cls}; f^{p,l}_{i,1}; f^{p,l}_{i,2}; \cdots; f^{p,l}_{i,N}] \tag{7}$$

$$F^s_{i,l} = E^l_s(F^s_{i,0}) = [f^{s,l}_{i,cls}; f^{s,l}_{i,1}; f^{s,l}_{i,2}; \cdots; f^{s,l}_{i,N}] \tag{8}$$

where E^l_p and E^l_s represent the first l Transformer layers of E_p and E_s, respectively. In the feature cross-transplantation process, we take local tokens in $F^p_{i,l}$ and $F^s_{i,l}$ as the objects to be transplanted, the non-cross remainder as the receptor, and the trained classifier W_p and W_s as the tester for whether the rejection occurs. If W_p and W_s can correctly classify the global features after transplantation, the transplanted features are deemed compatible with the original features. To achieve this challenging goal, the transplanted features must have consistent information to the original ones, i.e., the modal discrepancy will not affect the classification results. Therefore, CCFL can effectively alleviate the impact of modal discrepancy on pedestrian identity matching.

In this process, we first randomly select the local token $f^{s,l}_{i,j}$ to be transplanted from the sketch feature $F^s_{i,l}$, and then search for the corresponding local token $f^{p,l}_{i,j'}$ from the pedestrian photo feature

$\boldsymbol{F}_{i,l}^{p}$, where $j^{'}$ can be obtained by:

$$j^{'} = \arg\min_{n} \|\boldsymbol{f}_{i,j}^{s,l} - \boldsymbol{f}_{i,n}^{p,l}\|_2^2, n = 1, 2, \cdots, N \tag{9}$$

The input features of the l-th Transformer layer after cross-transplantation are respectively represented as:

$$\tilde{\boldsymbol{F}}_{i,l}^{p} = [\boldsymbol{f}_{i,cls}^{p,l}; \boldsymbol{f}_{i,1}^{p,l}; \boldsymbol{f}_{i,2}^{p,l}; \cdots; \boldsymbol{f}_{i,j}^{s,l}; \cdots; \boldsymbol{f}_{i,N}^{p,l}] \tag{10}$$

$$\tilde{\boldsymbol{F}}_{i,l}^{s} = [\boldsymbol{f}_{i,cls}^{s,l}; \boldsymbol{f}_{i,1}^{s,l}; \boldsymbol{f}_{i,2}^{s,l}; \cdots; \boldsymbol{f}_{i,j^{'}}^{p,l}; \cdots; \boldsymbol{f}_{i,N}^{s,l}] \tag{11}$$

The output features from each Transformer layer are cross transplanted as above and then sent to the next Transformer layer to extract features, and the final output class tokens of $\tilde{\boldsymbol{x}}_i^p$ and $\tilde{\boldsymbol{x}}_i^s$ are denoted as $\tilde{\boldsymbol{f}}_{i,cls}^{p}$ and $\tilde{\boldsymbol{f}}_{i,cls}^{s}$, respectively. In order to make $\boldsymbol{f}_{i,j^{'}}^{s,l}$ compatible with the features $[\boldsymbol{f}_{i,cls}^{p,l}; \boldsymbol{f}_{i,1}^{p,l}; \boldsymbol{f}_{i,2}^{p,l}; \cdots; \boldsymbol{f}_{i,N}^{p,l}]$ after removing $\boldsymbol{f}_{i,j}^{p,l}$, i.e., the classifier $\boldsymbol{W}_p$ and $\boldsymbol{W}_s$ are still able to correctly classify the final features $\tilde{\boldsymbol{f}}_{i,cls}^{p}$ and $\tilde{\boldsymbol{f}}_{i,cls}^{s}$, we use cross-entropy loss to further update the encoder $\boldsymbol{E}_p$ and $\boldsymbol{E}_s$:

$$\mathcal{L}_{id}^{p}(\boldsymbol{E}_p) = \sum_{i=1}^{B} -\boldsymbol{y}_i^p log(\boldsymbol{W}_p(\tilde{\boldsymbol{f}}_{i,cls}^{p})) \tag{12}$$

$$\mathcal{L}_{id}^{s}(\boldsymbol{E}_s) = \sum_{i=1}^{B} -\boldsymbol{y}_i^s log(\boldsymbol{W}_s(\tilde{\boldsymbol{f}}_{i,cls}^{s})) \tag{13}$$

3.3 Semantic Consistent Feature Construction

The non-corresponding information from the same location of sketches and pedestrian photos will lead to semantic information inconsistency in the learned features, affecting feature discrimination. To solve this problem, we propose SCFC, as shown in Figure 4.

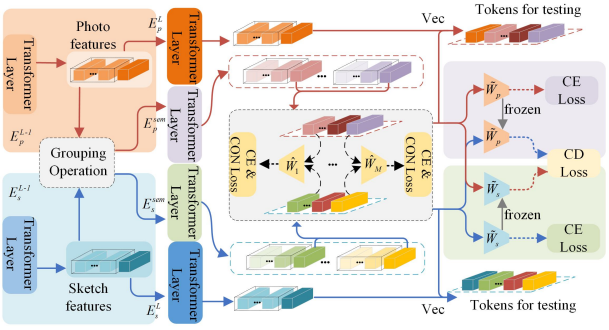

Figure 4: Semantic consistent feature construction.

In the proposed method, we first group the local tokens output by the Transformer layer and send them into the later Transformer layer to obtain the class tokens of each group. Each intra-modal set of class tokens is assigned different semantic information, and the corresponding inter-modal class tokens are given the same semantic information. Then, the class tokens assigned semantic information are used to construct a new global class token with semantic consistent for each modal image. Because the class tokens of different groups have class differences, it will guide the feature extraction network to extract features from different areas of pedestrian images, which helps to improve the representation ability of features.

Let $[\boldsymbol{f}_{i,cls}^{p,k}; \boldsymbol{f}_{i,1}^{p,k}; \boldsymbol{f}_{i,2}^{p,k}; \cdots; \boldsymbol{f}_{i,N}^{p,k}]$ and $[\boldsymbol{f}_{i,cls}^{s,k}; \boldsymbol{f}_{i,1}^{s,k}; \boldsymbol{f}_{i,2}^{s,k}; \cdots; \boldsymbol{f}_{i,N}^{s,k}]$ denote the outputs of the k-th Transformer layer, respectively. Both of them are divided into M groups. Each group consists of one class token and the same number of local tokens. The features of these groups are sent to the Transformer layer $\boldsymbol{E}_p^{sem}$ and $\boldsymbol{E}_s^{sem}$ (see Figure 4) separately to obtain the class tokens of each group $[\boldsymbol{f}_{i,cls,1}^{p}; \boldsymbol{f}_{i,cls,2}^{p}; \cdots; \boldsymbol{f}_{i,cls,M}^{p}]$ and $[\boldsymbol{f}_{i,cls,1}^{s}; \boldsymbol{f}_{i,cls,2}^{s}; \cdots; \boldsymbol{f}_{i,cls,M}^{s}]$. In order to enable each group of class tokens to express a semantic category while also representing the pedestrian identity information, we optimize the network parameters by using the loss functions:

$$\mathcal{L}_{id}^{p}(\bar{\boldsymbol{E}}_p, \boldsymbol{E}_p^{sem}, \hat{\boldsymbol{W}}_m) = \sum_{i=1}^{B} -\boldsymbol{y}_i^p log(\hat{\boldsymbol{W}}_m(\boldsymbol{f}_{i,cls,m}^{p})) \tag{14}$$

$$\mathcal{L}_{id}^{s}(\bar{\boldsymbol{E}}_s, \boldsymbol{E}_s^{sem}, \hat{\boldsymbol{W}}_m) = \sum_{i=1}^{B} -\boldsymbol{y}_i^s log(\hat{\boldsymbol{W}}_m(\boldsymbol{f}_{i,cls,m}^{s})) \tag{15}$$

$$\mathcal{L}_{id}^{p}(\bar{\boldsymbol{E}}_p, \boldsymbol{E}_p^{sem}, \tilde{\boldsymbol{W}}_p) = \sum_{i=1}^{B}\sum_{m=1}^{M} -\boldsymbol{y}_m^p log(\tilde{\boldsymbol{W}}_p(\boldsymbol{f}_{i,cls,m}^{p})) \tag{16}$$

$$\mathcal{L}_{id}^{s}(\bar{\boldsymbol{E}}_s, \boldsymbol{E}_s^{sem}, \tilde{\boldsymbol{W}}_s) = \sum_{i=1}^{B}\sum_{m=1}^{M} -\boldsymbol{y}_m^s log(\tilde{\boldsymbol{W}}_s(\boldsymbol{f}_{i,cls,m}^{s})) \tag{17}$$

where $\bar{\boldsymbol{E}}_p$ and $\bar{\boldsymbol{E}}_s$ represent the feature extraction networks composed of the first k Transformer layers of $\boldsymbol{E}_p$ and $\boldsymbol{E}_s$, respectively. $\hat{\boldsymbol{W}}_m$ $(m = 1, 2, \cdots M)$ is the shared pedestrian identity classifier of $\boldsymbol{f}_{i,cls,m}^{p}$ and $\boldsymbol{f}_{i,cls,m}^{s}$. $\boldsymbol{y}_i^p$ and $\boldsymbol{y}_i^s$ are the corresponding pedestrian identity labels of $\boldsymbol{f}_{i,cls,m}^{p}$ and $\boldsymbol{f}_{i,cls,m}^{s}$. $\tilde{\boldsymbol{W}}_p$ and $\tilde{\boldsymbol{W}}_s$ are the corresponding semantic category classifiers of $\boldsymbol{f}_{i,cls,m}^{p}$ and $\boldsymbol{f}_{i,cls,m}^{s}$. $\boldsymbol{y}_m^p$ and $\boldsymbol{y}_m^s$ are the predefined category labels, respectively.

Minimizing $\mathcal{L}_{id}^{p}(\bar{\boldsymbol{E}}_p, \boldsymbol{E}_p^{sem}, \hat{\boldsymbol{W}}_m)$ and $\mathcal{L}_{id}^{s}(\bar{\boldsymbol{E}}_s, \boldsymbol{E}_s^{sem}, \hat{\boldsymbol{W}}_m)$ only urges the network to focus on different parts of the pedestrian but cannot achieve feature semantic consistency. Thus, we assign the same semantic category information to $\boldsymbol{f}_{i,cls,m}^{p}$ and $\boldsymbol{f}_{i,cls,m}^{s}$. They are then sorted and combined according to semantic category information to build the semantic consistent feature representation. If $\boldsymbol{f}_{i,cls,m}^{p}$ and $\boldsymbol{f}_{i,cls,m}^{s}$ have the same category information, then $\boldsymbol{y}_m^p$ and $\boldsymbol{y}_m^s$ are also consistent, i.e., $\boldsymbol{y}_m^p = \boldsymbol{y}_m^s$. In this case, the classification result of feeding $\boldsymbol{f}_{i,cls,m}^{p}(\boldsymbol{f}_{i,cls,m}^{s})$ to $\tilde{\boldsymbol{W}}_s(\tilde{\boldsymbol{W}}_p)$ should be consistent with the one that feeding $\boldsymbol{f}_{i,cls,m}^{s}(\boldsymbol{f}_{i,cls,m}^{p})$ to $\tilde{\boldsymbol{W}}_s(\tilde{\boldsymbol{W}}_p)$. To this end, we introduce the consistency loss:

$$\mathcal{L}_{con}(\bar{\boldsymbol{E}}_p, \bar{\boldsymbol{E}}_s, \boldsymbol{E}_p^{sem}, \boldsymbol{E}_s^{sem}) = \sum_{i=1}^{B}\sum_{m=1}^{M} \|\boldsymbol{f}_{i,cls,m}^{s} - \boldsymbol{f}_{i,cls,m}^{p}\|_2^2 \tag{18}$$

and the cross discrimination loss:

$$\begin{aligned}&\mathcal{L}_{cd}(\bar{\boldsymbol{E}}_p, \bar{\boldsymbol{E}}_s, \boldsymbol{E}_p^{sem}, \boldsymbol{E}_s^{sem}) \\ &= -\sum_{i=1}^{B}\sum_{m=1}^{M} \left([\boldsymbol{y}_m^p log(\tilde{\boldsymbol{W}}_p(\boldsymbol{f}_{i,cls,m}^{s})) + \boldsymbol{y}_m^s log(\tilde{\boldsymbol{W}}_s(\boldsymbol{f}_{i,cls,m}^{s}))]\right. \\ &\left. + [\boldsymbol{y}_m^s log(\tilde{\boldsymbol{W}}_s(\boldsymbol{f}_{i,cls,m}^{p})) + \boldsymbol{y}_m^p log(\tilde{\boldsymbol{W}}_p(\boldsymbol{f}_{i,cls,m}^{p}))]\right)\end{aligned} \tag{19}$$

where $\boldsymbol{y}_m^p = \boldsymbol{y}_m^s$.

In the above process, if $f^p_{i,cls,m}$ and $f^s_{i,cls,m}$ have different semantic information, the classification results will be different by feeding $f^p_{i,cls,m}$ and $f^s_{i,cls,m}$ into the same classifier $\tilde{W}_p$ or $\tilde{W}_s$. Thus, minimizing $\mathcal{L}_{cd}(\bar{E}_p, \bar{E}_s, E^{sem}_p, E^{sem}_s)$ will effectively facilitate feature extraction network to extract semantic consistent features $f^p_{i,cls,m}$ and $f^s_{i,cls,m}$. Meanwhile, the proposed SCFC can further narrow the modal gap between sketches and photos. It is mainly because if $f^p_{i,cls,m}$ and $f^s_{i,cls,m}$ carry modal information that affects the classification results, then the classification results will be inconsistent as they are fed to classifier $\tilde{W}_p$ or $\tilde{W}_s$. This is contradict to the objective of the loss function (19).

After obtaining the semantic consistent features $f^p_{i,cls,m}$ and $f^s_{i,cls,m}$, we combine the class tokens of each group with the global class token to form a new global class token for pedestrian identity matching:

$$\bar{f}^p_{i,cls} = vec(f^p_{i,cls,1}, f^p_{i,cls,2}, \cdots, f^p_{i,cls,M}, f^p_{i,cls}) \tag{20}$$

$$\bar{f}^s_{i,cls} = vec(f^s_{i,cls,1}, f^s_{i,cls,2}, \cdots, f^s_{i,cls,M}, f^s_{i,cls}) \tag{21}$$

where $vec(\cdot)$ represents a vectorization operation. $f^p_{i,cls}$ and $f^s_{i,cls}$ are the corresponding global class tokens of pedestrian photos and sketches, respectively. Finally, the total loss function is formulated as:

$$\begin{aligned}
&\mathcal{L}(E_p, E_s, E^{sem}_p, E^{sem}_s, W_p, W_s, \tilde{W}_p, \tilde{W}_s, \hat{W}_m) \\
&= [\mathcal{L}^p_{id}(E_p, W_p) + \mathcal{L}^s_{id}(E_s, W_s)] + [\mathcal{L}^p_{tri}(E_p) + \mathcal{L}^s_{tri}(E_s)] \\
&+ [\mathcal{L}^p_{id}(E_p) + \mathcal{L}^s_{id}(E_s)] + [\mathcal{L}^p_{id}(\bar{E}_p, E^{sem}_p, \hat{W}_p) \\
&+ \mathcal{L}^s_{id}(\bar{E}_s, \bar{E}^{sem}_s, \hat{W}_s)] + [\mathcal{L}^p_{id}(\bar{E}_p, E^{sem}_p, \tilde{W}_m) \\
&+ \mathcal{L}^s_{id}(\bar{E}_s, E^{sem}_s, \tilde{W}_m)] + \mathcal{L}_{con}(\bar{E}_p, \bar{E}_s, E^{sem}_p, E^{sem}_s) \\
&+ \mathcal{L}_{cd}(\bar{E}_p, \bar{E}_s, E^{sem}_p, E^{sem}_s)
\end{aligned} \tag{22}$$

4 EXPERIMENTS

4.1 Datasets

Currently, there is only one public Sketch Re-ID dataset Sketch Re-ID [18]. The dataset contains 200 persons, each with one sketch and two photos. According to the experimental setup in [18], 150 persons are used for training and other 50 persons for testing. In addition, we also validate the proposed model on two publicly available datasets QMUL-Shoe-v2 [30] and QMUL-Chair-v2 [30] that are commonly used for fine-grained sketch image retrieval tasks. QMUL-Shoe-v2 includes 6730 sketches and 2000 photos, one ID for each photo. Following previous work [16], we use 1800 sketches and photos for training and testing on the remaining samples. QMUL-Chair-V2 has a total of 1275 sketches and 400 photos, with one ID for each photo. This paper uses samples of 300 IDs as the training set and the remaining samples as the test set.

4.2 Implementation Details

The proposed model is implemented under the Pytorch framework [19]. All experiments are conducted on an NVIDIA GeForce RTX 3090 24GB GPU. We use the Transformer model pre-trained on ImageNet [7] as the backbone. Before the images are input to the feature extraction network, all images are resized to 256×256, the block size is set to 16×16, and the sliding step S is 16. Besides, we use horizontal flipping, padding, random cropping, random erasing, etc. in [32] for data augmentation. In the experiments, the batchsize of each modality is set to 16. 8 IDs are included in a batch, and two images are selected for each ID. According to [10], we set the total number of Transformer layers L to 12. The initial learning rate is set to 0.0015 for the SGD optimizer, and the weight decay of the optimizer is set to 10^{-4}. Then, the cosine learning rate decay strategy is used to adjust the learning rate throughout the training process. The model is trained with a total of 100 epochs. To reduce the complexity, the semantic consistent feature construction only occurs at the penultimate layer of feature extraction network E_p and E_s, i.e., $k = L - 1$ (see Figure 2 for details). The Euclidean Distance (ED) is used to match pedestrian identities, and the cumulative match characteristic (CMC)[26] and mean average precision (mAP)[31] are employed to evaluate model performance.

4.3 Comparison with State-of-the-arts

In this section, we compare the proposed model with the existing methods Dense-HOG+rankSVM [11, 18], TripetSN [29], GNSiamese [20], AFLNet [18] and LMDF [9] on Sketch Re-ID task to verify the superiority of the proposed method. As shown in Table 1, the proposed method significantly outperforms the state-of-the-art methods compared. Specifically, the proposed method reaches 86.0% Rank-1 accuracy, which is 37.0% higher than the sub-optimal method LMDF. This is because we adopt the idea of compatible transplantation and design a cross-compatible module, which effectively reduces the gap between modalities and prevents the loss of discriminative features, promoting model's generalization ability. In addition, the feature semantic consistency effectively avoids inconsistency of information located at the same location and the fact that the model tends to extract features from the most discriminative local region, which is not fully considered in the comparison methods, resulting in weak recognition performance.

Table 1 Performance comparisons on Sketch Re-ID dataset.

Methods	Rank-1	Rank-5	Rank-10	mAP
Dense-HOG+rankSVM[11, 18]	5.1%	16.8%	28.3%	–
Triplet SN[18, 29]	9.0%	26.8%	42.2%	–
GNSiamese[18, 20]	28.9%	54.0%	62.4%	–
AFLNet[18]	34.0%	56.3%	72.5%	–
LMDF[9]	49.0%	70.4%	80.2%	–
Proposed	**86.0%**	**98.0%**	**100.0%**	**83.7%**

In order to evaluate the generalization of the proposed method, we verify the performance of the proposed method on the fine-grained sketch image retrieval task. In this process, two commonly used datasets QMUL-Shoe-v2 and QMUL-Chair-v2 are used to evaluate model performance. As shown in Table 2, the proposed method surpasses all comparison methods on all evaluation metrics. Specifically, on QMUL-Chair-v2, the proposed method reaches 74.3% rank-1 recognition rate, exceeding the sub-optimal algorithm SketchAA-Graph by 21.4%. On QMUL-Shoe-v2, the recognition rate of the proposed method on Rank-1 reaches 33.5%, exceeding the sub-optimal algorithm by 1.2%. As such, the proposed algorithm has better generalization ability.

Table 2 Performance comparisons on QMUL-Chair-V2 and QMUL-Shoe-V2.

Methods	QMUL-Chair-V2			QMUL-Shoe-V2		
	Rank-1	Rank-10	mAP	Rank-1	Rank-10	mAP
TripletSN[1, 29]	47.4%	84.3%	–	28.7%	71.6%	–
SN-HOLEF[1, 23]	50.7%	86.3%	–	31.2%	74.6%	–
SN-RL [1, 2]	51.2%	86.9%	–	30.8%	74.2%	–
Siamese-Tri-SA[22, 28]	47.2%	90.9%	–	31.1%	80.0%	–
SketchAA-Graph[28]	52.9%	94.9%	–	32.3%	79.6%	–
Proposed	**74.3%**	**97.4%**	**83.3%**	**33.5%**	**80.2%**	**48.7%**

4.4 Ablation Study

The proposed method mainly consists of CCFL and SCFC. To validate the contribution of these two modules to the proposed model, we experimentally verify the role of each module. In this process, we use the Transformer network trained with identity loss and triplet loss as Baseline. Baseline+CCFL denotes a baseline model added cross-compatible feature learning module. Baseline+SCFC means a baseline model added semantic consistent feature construction, while Baseline+CCFL+SCFC means the full model.

Effectiveness of CCFL: To verify the effectiveness of the CCFL, we compare Baseline+CCFL and Baseline+CCFL+SCFC to Baseline and Baseline+SCFC, respectively. As can be seen from Tables 3 and 4, the CCFL module can effectively improve the performance of the model. Specifically, on Sketch Re-ID task, Baseline increases Rank-1 recognition accuracy from 66.0% to 76.0% after adding CCFL module, while Baseline+SCFC also has a significant performance improvement with CCFL. Meanwhile, on other datasets based on sketch image retrieval (as shown in Table 4), CCFL shows significant validity as well.

Effectiveness of SCFC: The SCFC is added to Baseline and Baseline+CCFL respectively to verify its effectiveness. As can be seen from the experimental results in Table 3, on Sketch Re-ID task, Baseline+SCFC increases the recognition accuracy on Rank-1 from 66.0% to 78.0% compared to Baseline. Compared with Baseline+CCFL, Baseline+CCFL+SCFC also provides an improvement. On other datasets based on sketch image retrieval (as shown in Table 4), the effectiveness of the SCFC is to be further verified.

Table 3 Ablation study of each module on Sketch Re-ID dataset.

Baseline	CCFL	SCFC	Rank-1	Rank-5	Rank-10	mAP
✓	✗	✗	66.0%	94.0%	96.0%	69.7%
✓	✓	✗	76.0%	94.0%	96.0%	77.0%
✓	✗	✓	78.0%	98.0%	100.0%	79.9%
✓	✓	✓	86.0%	98.0%	100.0%	83.7%

Table 4 Ablation study of each module on QMUL-Shoe-v2.

Baseline	CCFL	SCFC	Rank-1	Rank-5	Rank-10	mAP
✓	✗	✗	23.6%	53.8%	67.3%	38.1%
✓	✓	✗	25.4%	59.2%	72.7%	41.0%
✓	✗	✓	31.5%	62.2%	76.0%	46.0%
✓	✓	✓	33.5%	65.5%	80.2%	48.7%

At the same time, the SCFC urges the network to pay attention to the information of various parts of the pedestrian. To verify this, Figure 5 shows the changes in the areas of interest to the network before and after adding SCFC. It can be seen that the network tends to extract features from the most discriminative areas of the pedestrian before the integration of the SCFC. After the integration, the network extracts discriminative features from different pedestrian parts, implying that the SCFC enables the network to avoid focusing on the most discriminative region of an image.

Figure 5: Impact of the SCFC on discriminant feature extraction. Images in the first row are the attention maps generated by Baseline+CCFL, and in the second row are the attention maps generated by Baseline+CCFL+SCFC. The warmer the color is, the more attention the region receives of the feature extraction network.

4.5 Discussion

In this section, we analyze the effects of the number of CT blocks in each layer of the cross-compatibility module, the effects of the number of groups in SCFC, and the contribution of classifiers $\tilde{W}_p$ and $\tilde{W}_s$ for pedestrian identity matching. Figure 6 shows the effect of the number of cross blocks on model performance on the Sketch Re-ID dataset and the sketch image retrieval dataset QMUL-Chair-V2.

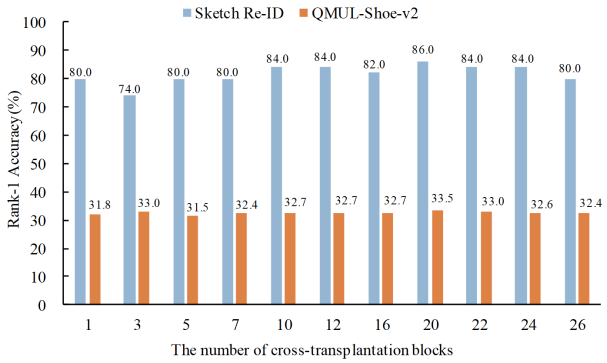

Figure 6: Influence of the number of transplanted blocks on Rank-1 in the cross-compatible feature learning.

In each Transformer layer, the number of CT blocks is set from 1 to 26. For the Sketch re-ID dataset and the sketch image retrieval dataset, the optimal model performance is achieved when the number of CT blocks is 20. It might because too many CT blocks could

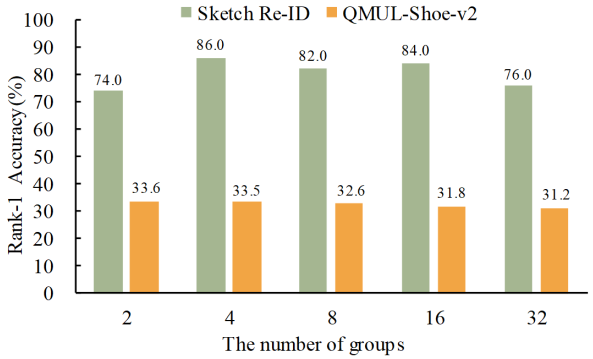

Figure 7: Influence of the number of groups on Rank-1 in the semantic consistent feature construction.

affect the network's ability to extract discriminative features from a single complete image, limiting further improvement in model performance.

Figure 7 describes the effect of the number of groups on recognition performance in SCFC. We can tell that the model performance increases with the number of groups from 2 to 4. The model performance decreases once the number of groups is greater than 4. Therefore, in this work, the local tokens of the penultimate Transformer layer are divided into 4 groups and sent to the last Transformer layer to build the semantic consistent features.

To this end, we introduce classifiers $\tilde{\boldsymbol{W}}_p$ and $\tilde{\boldsymbol{W}}_s$ to the SCFC module and use them crosswise to facilitate the network to assign the same semantic information to the corresponding group of the class tokens. Moreover, the cross-use of the two classifiers can further promote the network to narrow the gap between different modalities. The experiment results are compared with those of classifiers $\tilde{\boldsymbol{W}}_p$ and $\tilde{\boldsymbol{W}}_s$ without cross (i.e., only one classifier is used). As can be seen from the results in Figure 8, the performance of the model deteriorates when $\tilde{\boldsymbol{W}}_p$ and $\tilde{\boldsymbol{W}}_s$ are not crossed, which confirms the contribution of cross-using $\tilde{\boldsymbol{W}}_p$ and $\tilde{\boldsymbol{W}}_s$.

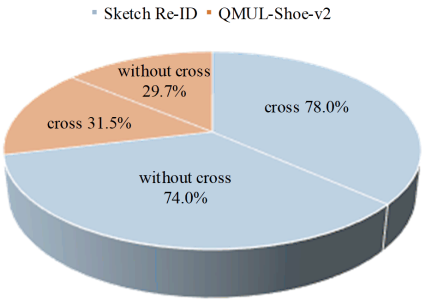

Figure 8: Effect of cross-using of classifiers $\tilde{W}_p$ and $\tilde{W}_s$ on Rank-1. "cross" means classifiers $\tilde{W}_p$ and $\tilde{W}_s$ are crossed in Baseline+SCFC. "without cross" means only single classifier is used in Baseline+SCFC.

5 CONCLUSION

In this work, a novel sketch-photo Re-ID method is proposed. It consists of a compatible feature learning mechanism based on CT to alleviate the modal gap between sketches and pedestrian photos by eliminating the modal discrepancy. To further improve the representation ability of features, the SCFC that recombines class tokens with different semantics in terms of semantic information as the final feature to describe the pedestrian is presented. The mechanism helps to achieve feature semantic consistency and narrows the gap between two modalities. Compared with the sate-of-the-art methods, the proposed model achieves better results on the Sketch Re-ID, QMUL-Shoe-v2 and QMUL-Chair-v2 datasets. The contributions of different modules are investigated by the ablation study. The results show that the proposed model is suitable for Sketch Re-ID and sketch-based image retrieval.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China under Grant (61966021, 62161015).

REFERENCES

[1] Ayan Kumar Bhunia, Pinaki Nath Chowdhury, Aneeshan Sain, Yongxin Yang, Tao Xiang, and Yi-Zhe Song. 2021. More photos are all you need: semi-supervised learning for fine-grained sketch based image retrieval. In *CVPR*. 4245–4254.
[2] Ayan Kumar Bhunia, Yongxin Yang, Timothy M. Hospedales, Tao Xiang, and Yi-Zhe Song. 2020. Sketch less for more: on-the-fly fine-grained sketch-based image retrieval. In *CVPR*. 9779–9788.
[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with transformers. In *ECCV*. 213–229.
[4] Chun-Fu (Richard) Chen, Quanfu Fan, and Rameswar Panda. 2021. CrossViT: Cross-Attention Multi-Scale Vision Transformer for Image Classification. In *ICCV*. 347–356.
[5] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. 2021. Pre-trained image processing transformer. In *CVPR*. 12299–12310.
[6] Yangdong Chen, Zhaolong Zhang, Yanfei Wang, Yuejie Zhang, Rui Feng, Tao Zhang, and Weiguo Fan. 2022. AE-Net: Fine-grained sketch-based image retrieval via attention-enhanced network. *Pattern Recognition* 122 (2022), 108291.
[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*. 248–255.
[8] Changxing Ding, Kan Wang, Pengfei Wang, and Dacheng Tao. 2022. Multi-task learning with coarse priors for robust part-aware person re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 3 (2022), 1474–1488.
[9] Shaojun Gui, Yu Zhu, Xiangxiang Qin, and Xiaofeng Ling. 2020. Learning multi-level domain invariant features for sketch re-identification. *Neurocomputing* 403 (2020), 294–303.
[10] Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. 2021. TransReID: transformer-based object re-identification. In *ICCV*.
[11] Rui Hu and John Collomosse. 2013. A performance evaluation of gradient field hog descriptor for sketch based image retrieval. *Computer Vision and Image Understanding* 117, 7 (2013), 790–806.
[12] Huafeng Li, Yiwen Chen, Dapeng Tao, Zhengtao Yu, and Guanqiu Qi. 2021. Attribute-aligned domain-invariant feature learning for unsupervised domain adaptation person re-identification. *IEEE Transactions on Information Forensics and Security* 16 (2021), 1480–1494.
[13] Yulin Li, Jianfeng He, Tianzhu Zhang, Xiang Liu, Yongdong Zhang, and Feng Wu. 2021. Diverse part discovery: occluded person re-identification with part-aware transformer. In *CVPR*. 2898–2907.
[14] Shengcai Liao and Ling Shao. 2021. Transformer-based deep image matching for generalizable person re-identification. *arXiv preprint arXiv:2105.14432* (2021).
[15] Hangyu Lin, Yanwei Fu, Peng Lu, Shaogang Gong, Xiangyang Xue, and Yugang Jiang. 2019. TC-Net for iSBIR: Triplet classification network for instance-level sketch based image retrieval. In *ACMMM*. 1676–1684.
[16] Zhipu Liu, Lei Zhang, and Yang Yang. 2020. Hierarchical bi-directional feature perception network for person re-identification. In *ACMMM*. 4289–4298.
[17] Zhongxing Ma, Yifan Zhao, and Jia Li. 2021. Pose-guided inter- and intra-part relational transformer for occluded person re-identification. In *ACMMM*. 1487–1496.

[18] Lu Pang, Yaowei Wang, Yi-Zhe Song, Tiejun Huang, and Yonghong Tian. 2018. Cross-domain adversarial feature learning for sketch re-identification. In *ACMMM*. 609–617.
[19] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *NeurIPS* 32, 8026–8037.
[20] Patsorn Sangkloy, Nathan Burnell, Cusuh Ham, and James Hays. 2016. The sketchy database: learning to retrieve badly drawn bunnies. *ACM Transactions on Graphics* 35, 4 (2016), 1–12.
[21] Jifei Song, Kaiyue Pang, Yi-Zhe Song, Tao Xiang, and Timothy M. Hospedales. 2018. Learning to sketch with shortcut cycle consistency. In *CVPR*. 801–810.
[22] Jifei Song, Qian Yu, Yi-Zhe Song, Tao Xiang, and Timothy M. Hospedales. 2017. Deep spatial-semantic attention for fine-grained sketch-based image retrieval. In *ICCV*. 5551–5560.
[23] Jifei Song, Yi zhe Song, Tony Xiang, and Timothy Hospedales. 2017. Fine-Grained Image Retrieval: the Text/Sketch Input Dilemma. In *BMVC*. 45.1–45.12.
[24] Peize Sun, Jinkun Cao, Yi Jiang, Rufeng Zhang, Enze Xie, Zehuan Yuan, Changhu Wang, and Ping Luo. 2020. Transtrack: multiple-object tracking with transformer. *arXiv preprint arXiv:2012.15460* (2020).
[25] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NeurIPS*. 5998–6008.
[26] Xiaogang Wang, Gianfranco Doretto, Thomas Sebastian, Jens Rittscher, and Peter Tu. 2007. Shape and appearance context modeling. In *ICCV*. 1–8.
[27] Enze Xie, Wenjia Wang, Wenhai Wang, Peize Sun, Hang Xu, Ding Liang, and Ping Luo. 2021. Segmenting transparent objects in the wild with transformer. In *IJCAI*. 1194–1200.
[28] Lan Yang, Kaiyue Pang, Honggang Zhang, and Yi-Zhe Song. 2021. SketchAA: abstract representation for abstract sketches. In *ICCV*. 10077–10086.
[29] Qian Yu, Feng Liu, Yi-Zhe Song, Tao Xiang, Timothy M. Hospedales, and Chen Change Loy. 2016. Sketch me that shoe. In *CVPR*. 799–807.
[30] Qian Yu, Jifei Song, Yi-Zhe Song, Tao Xiang, and Timothy M. Hospedales. 2021. Fine-grained instance-level sketch-based image retrieval. *International Journal of Computer Vision* 129 (2021), 484–500.
[31] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. 2015. Scalable person re-identification: A benchmark. In *ICCV*. 1116–1124.
[32] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. 2020. Random erasing data augmentation. In *AAAI*. 13001–13008.
[33] Kuan Zhu, Haiyun Guo, Shiliang Zhang, Yaowei Wang, Gaopan Huang, Honglin Qiao, Jing Liu, Jinqiao Wang, and Ming Tang. 2021. AAformer: auto-aligned transformer for person re-identification. *arXiv preprint arXiv:2104.00921* (2021).