Differentiable Auxiliary Learning for Sketch Re-Identification

Xingyu Liu[1], Xu Cheng[1*], Haoyu Chen[2], Hao Yu[1,2], Guoying Zhao[2]

[1]School of Computer Science, Nanjing University of Information Science and Technology, China
[2]Center for Machine Vision and Signal Analysis, University of Oulu, Finland
{xingyu, xcheng, yuhao}@nuist.edu.cn, {chen.haoyu, guoying.zhao}@oulu.fi

Abstract

Sketch re-identification (Re-ID) seeks to match pedestrians' photos from surveillance videos with corresponding sketches. However, we observe that existing works still have two critical limitations: (i) cross- and intra-modality discrepancies hinder the extraction of modality-shared features, (ii) standard triplet loss fails to constrain latent feature distribution in each modality with inadequate samples. To overcome the above issues, we propose a differentiable auxiliary learning network (DALNet) to explore a robust auxiliary modality for Sketch Re-ID. Specifically, for (i) we construct an auxiliary modality by using a dynamic auxiliary generator (DAG) to bridge the gap between sketch and photo modalities. The auxiliary modality highlights the described person in photos to mitigate background clutter and learns sketch style through style refinement. Moreover, a modality interactive attention module (MIA) is presented to align the features and learn the invariant patterns of two modalities by auxiliary modality. To address (ii), we propose a multi-modality collaborative learning scheme (MMCL) to align the latent distribution of three modalities. An intra-modality circle loss in MMCL brings learned global and modality-shared features of the same identity closer in the case of insufficient samples within each modality. Extensive experiments verify the superior performance of our DALNet over the state-of-the-art methods for Sketch Re-ID, and the generalization in sketch-based image retrieval and sketch-photo face recognition tasks.

Introduction

Person re-identification (Re-ID) (Ye et al. 2021) is commonly employed for matching pedestrian images captured by multiple non-overlapping cameras. However, traditional Re-ID methods can hardly be used when there are only the descriptions of the corresponding individuals in the absence of the query photos. To address this situation, Sketch Re-ID (Pang et al. 2018) has emerged to match the sketches, drawn by experts according to witness descriptions, with pedestrian photos captured by surveillance cameras.

Due to the different presentation intrinsicality between sketches and photos, Sketch Re-ID suffers from undesired cross-modality differences. As shown in Figure 1(a), photos from different camera views can capture more visual information (e.g., color, pose, background, and lighting), while hand-drawn sketches from a frontal view are very abstract with only contour information and prominent features. On account of diverse painting styles, noticeable differences in the portrayal of characters emerge within the sketch modality. Besides, background noise and light variety within the photo modality pose a challenge for aligning sketches and photos of the same identity.

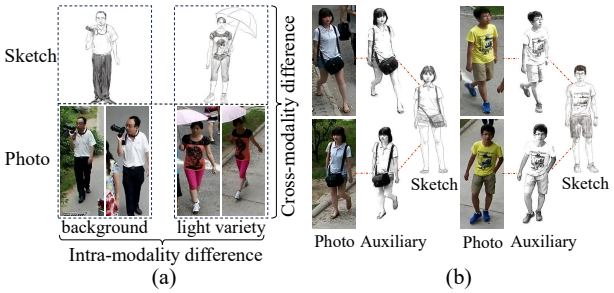


Figure 1: (a) Sketch-photo sample pairs from the PKU-Sketch dataset. (b) Visualization of our auxiliary images from DAG. Our auxiliary modality possesses a noise-free structure similar to photos and a sketch-like texture, thereby facilitating the learning of modality-shared features.

Existing Sketch Re-ID methods (Pang et al. 2018; Gui et al. 2020) rely on extracting shared patterns and standard triplet loss for cross-modality metric optimization. Typically, a two-stream network (Zhu et al. 2022) is utilized to extract features across different modalities, and these features are mapped into a common feature space. Despite the encouraging achievements, these methods still struggle with huge modality discrepancies. Recently, Zhang *et al*. (Zhang et al. 2022) developed cross-compatible embedding and feature construction to enhance feature discrimination among different modalities. However, cross transplantation inevitably introduces extra noise into the learned patterns due to local exchange between two modalities. Apart from enhancing feature discrimination, bridging the modality gap is noteworthy in Sketch Re-ID. An auxiliary sketch modality was proposed to guide the asymmetrical disentanglement in the transformer framework (Chen et al. 2022). But they

*Corresponding author (Email: xcheng@nuist.edu.cn).

migrated the photo-to-sketch synthesis method (Xiang et al. 2022) into Sketch Re-ID, producing much sketch-like background clutter in cross-modality retrieval. The aforementioned methods aim to address cross-modality differences but neglect significant variations within each modality.

We cast the Sketch Re-ID as a distribution matching problem, where the goal is to train a model such that the learned distribution of latent representations is invariant across different modalities and within each modality. To this end, we propose a differentiable auxiliary learning network (DALNet) for Sketch Re-ID. Initially, the dynamic auxiliary generator (DAG) generates the auxiliary modality from photo modality under the constraint of style refinement learning (SRL), setting the stage for learning distribution matching. To overcome intra-modality variations, the auxiliary modality highlights the pertinent patterns of the human body while intentionally mitigating any potential interference caused by background clutter and light variations in photos during the training process. Also, the auxiliary modality can be dynamically adjusted to approximate the specific sketch style by SRL strategy, alleviating the impact of diverse painting styles. Figure 1(b) illustrates the visualization of our auxiliary modality, which not only has a similar style to the hand-drawn sketches but also maintains the same structure as the photos without background noise.

To learn the invariant representations across photo and sketch modalities, we design a modality interactive attention module (MIA). MIA refines the semantic information of sketch and photo modalities by utilizing auxiliary modality. It serves as a bridge to achieve multi-modality fine-grained feature interaction. Additionally, we design a multi-modality collaborative learning scheme (MMCL) instead of modality replacement (Bhunia et al. 2021) to learn robust feature relationships across photo, sketch and auxiliary modalities at the identity and distribution levels. Unlike the previous methods working around triplet loss for cross-modality metric, we narrow the gap between sketch and photo modalities while considering the sub-optimal latent space issues caused by the intra-modality discrepancy. An intra-modality circle loss ($\mathcal{L}_{IM}$) in MMCL is designed to bring the features of the same identity closer within each modality. For the scarcity of samples, $\mathcal{L}_{IM}$ leverages learned global features and modality-shared features of the same sample at different stages. These features are the most dissimilar in the same sample, increasing representation richness and effectively alleviating feature latent distribution imbalances. In general, our contributions are summarized as follows.

- We propose a differentiable auxiliary learning network with a powerful auxiliary modality to tackle the issues of cross- and intra-modality discrepancies, and feature distribution with insufficient samples.
- A lightweight dynamic auxiliary generator is designed to generate the auxiliary modality, which is updated to highlight the described person in photos and learn sketch style with style refinement learning.
- We present a modality interactive attention module to align the features and learn the invariant patterns of photo and sketch modalities by auxiliary modality. Moreover, a multi-modality collaborative learning scheme is designed to align the latent distribution of three modalities, even in the case of limited samples.
- Extensive experiments verify the superior performance of our DALNet over state-of-the-art methods for Sketch Re-ID and other related tasks, such as sketch-based image retrieval and sketch-photo face recognition.

Related Work

Sketch-Based Image Retrieval

Sketch-based image retrieval (SBIR) utilizes hand-drawn sketches as the query set to retrieve relevant photos from a database. Existing category-level SBIR methods (Bui et al. 2017; Collomosse, Bui, and Jin 2019) make use of Siamese networks with ranking losses to learn a joint embedding space. In addition, Li *et al*. (Li et al. 2014) first developed deformable part-based models for fine-grained SBIR, which aims at instance specific sketch-photo matching. Yu *et al*. (Yu et al. 2016) introduced a new database containing two categories with deep triplet ranking annotations and explored a shared feature space for sketch and photo modalities. After that, some methods are developed to improve feature representation by using attention mechanisms (Song et al. 2018; Sain et al. 2020; Yu et al. 2021), generative learning (Pang et al. 2017; Chen et al. 2021), reinforcement learning (Bhunia et al. 2020), data augmentation (Bhunia et al. 2021) and self-supervised pre-training (Pang et al. 2020). In addition, some methods utilized feature-level content-style disentanglement (Deng et al. 2020; Sain et al. 2021) to decompose features into different representations.

Sketch Re-Identification

Sketch Re-ID is one of the tasks in sketch retrieval, which aims to match a hand-drawn sketch with a relevant photo of a person. Pang *et al*. (Pang et al. 2018) firstly introduced a new sketch-photo dataset with cross-domain annotations and utilized cross-domain adversarial feature learning to narrow the gap between sketches and photos. Later, a novel approach for Sketch Re-ID (Gui et al. 2020) emerged, learning multi-level domain invariant features. Chen *et al*. (Chen et al. 2022) proposed an asymmetrical disentanglement and dynamic synthesis method to handle modality discrepancy in the transformer framework. Zhang *et al*. (Zhang et al. 2022) proposed a cross-compatible learning mechanism and introduced a semantic consistent feature construction mechanism to handle non-corresponding information between two modalities. Nevertheless, the above-mentioned methods still suffer from greater challenges such as modality discrepancy, background clutter, style diversity, etc. To overcome these issues, we cast the Sketch Re-ID as a distribution matching problem. The auxiliary modality is utilized to learn the invariant representations across photo and sketch modalities, and align the distribution over cross- and intra-modality features, even with limited samples in each modality.

Methodology

In this section, we introduce the differentiable auxiliary learning network (DALNet) for Sketch Re-ID, which can

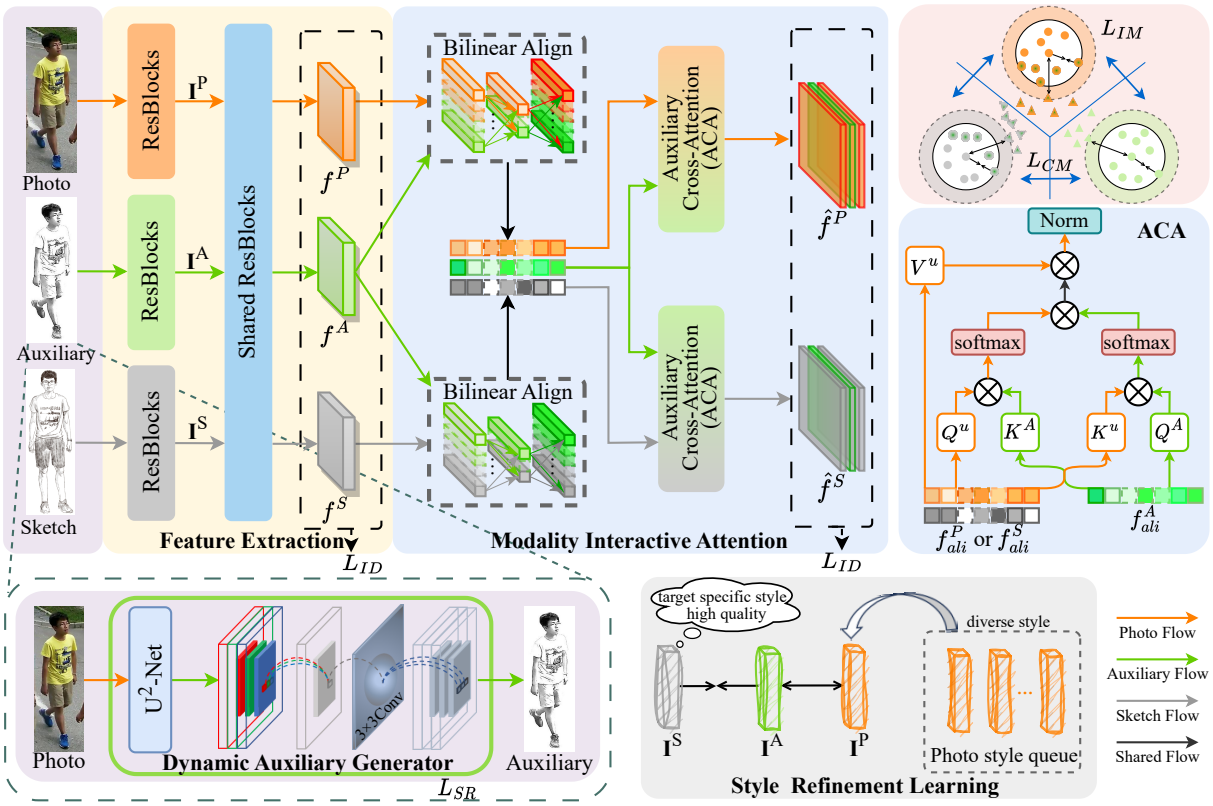


Figure 2: The overall architecture of the proposed method for Sketch Re-ID. During the inference, we do not utilize the dynamic auxiliary generator, and only photo and sketch modalities are utilized to perform cross-modality retrieval.

adaptively refine the generation of auxiliary modality and learn cross-modality latent representations. The overview of our DALNet is shown in Figure 2 and more details are discussed in the following subsections.

Overview

Let $P = \{x_i^P\}_{i=1}^{N_P}$, $S = \{x_i^S\}_{i=1}^{N_S}$ and $A = \{x_i^A\}_{i=1}^{N_A}$ denote photo, sketch and auxiliary modalities, respectively. N_P, N_S and N_A denote the number of samples. First, a sketch-like auxiliary modality is generated from the photo modality by using the proposed dynamic auxiliary generator (DAG). The parameters of the generator are dynamically adjusted under the constraint of style refinement learning. Then, we adopt a three-stream network as the feature extractor. The first three ResBlocks are utilized to extract photo, auxiliary and sketch features, respectively. After that, the learned global features of three modalities are fed into weight shared ResBlocks to learn the high-level features. Further, the modality-shared features of three modalities are input into modality interactive attention module (MIA), which consists of a bilinear align module (BAM) and a auxiliary cross-attention module (ACA). BAM is used to align the features and ACA is employed to learn the invariant patterns of photo and sketch modalities. In addition, the effective cross-modality circle loss ($\mathcal{L}_{CM}$) and intra-modality circle loss ($\mathcal{L}_{IM}$) are presented to constrain the feature distribution of three modalities.

Dynamic Auxiliary Generator

Before learning the invariant latent patterns between photo and sketch modalities, we propose a dynamic auxiliary generator (DAG) to generate a robust auxiliary modality. As shown in Figure 2, the proposed DAG consists of U^2-Net (Qin et al. 2020), 1×1 convolutional layer, ReLU activation layer and 3×3 convolutional layer. Photos are regarded as the input of the generator. First, the pre-trained U^2-Net model is utilized to separate the foreground from the background in an image, reducing irrelevant background noise. The 1×1 convolutional layer transforms three-channel photos into one-channel grayscale photos. A ReLU activation layer is provided to improve the non-linear representation capability. Then, a 3×3 convolutional layer is exploited to detect object edges in the photos. Its convolutional kernel weight matrix is designed as an edge detection filter to enhance edge information, and weights are dynamically updated during the training process. Finally, one-channel arrays are converted into three-channel sketch-like auxiliary modality through channel replication.

Based on the above-mentioned DAG structure, which siphons off knowledge from the photos and is similar to sketches, the auxiliary modality is generated. Compared with the existing auxiliary modality generator used for

Sketch Re-ID (Chen et al. 2022) that introduced obvious background noise from photos, the proposed lightweight DAG generates high-quality auxiliary modality without background clutter and light interference.

Style Refinement Learning

Due to the distinct sketch styles in several tasks such as Sketch Re-ID, sketch-based image retrieval, and sketch-photo face recognition, the generator with fixed parameters fails to describe the entire structure of the object, thereby impacting the recognition accuracy. To improve the flexibility and generalization ability of the generator, we propose a style refinement learning strategy to refine the generator. In our settings, the hand-drawn sketches are considered to have a consistent style with the generated auxiliary modality, while the style of the photo is inconsistent with that of the auxiliary modality. We exploit three ResBlocks, Generalized-Mean (GeM) pooling (Radenović, Tolias, and Chum 2018) and a classifier to obtain the style texture information $\mathbf{I}$ for different modalities.

Inspired by the idea of contrastive learning (Chen et al. 2020), a style refinement loss $\mathcal{L}_{SR}$ is designed by:

$$\mathcal{L}_{SR} = -\log \frac{\exp(\frac{(\mathbf{I}^A)^T \mathbf{I}^S}{\xi})}{\exp(\frac{(\mathbf{I}^A)^T \mathbf{I}^S}{\xi}) + \sum_{i=1}^{N} \exp(\frac{(\mathbf{I}^A)^T \mathbf{I}_i^P}{\xi})}, \quad (1)$$

where ξ denotes a temperature parameter; N denotes the number of random photo samples in the current and preceding mini-batches; $\mathbf{I^P}$, $\mathbf{I^A}$ and $\mathbf{I^S}$ stand for photo, auxiliary and sketch style information, respectively.

Modality Interactive Attention

The unaligned patterns in the latent distribution between photo and sketch modalities are one of the key factors affecting pedestrian identity matching. The reason is that sketches usually only contain simple contour information about the human body, while photos provide more visual information. To address this issue, we present a modality interactive attention module (MIA) to learn modality-invariant features. MIA is composed of a bilinear align module (BAM) and an auxiliary cross-attention module (ACA). Auxiliary modality features play a guiding role in highlighting the common parts of the sketches and photos.

First, the learned global features from the three-stream network are fed into BAM to align the features of photo and sketch modalities by auxiliary modality. Specifically, the photo feature $\boldsymbol{f}^P$ and sketch feature $\boldsymbol{f}^S$ are first paired with auxiliary feature $\boldsymbol{f}^A$ and input to the BAM, which consists of two linear layers and a ReLU activation. The first linear layer is utilized to compress the feature dimension of the channel to a quarter and the second one can map the intermediate representation back to the original dimension, which not only enhances the feature correlation of two modalities but also reduces the number of network parameters.

After that, the similarity score $\boldsymbol{S}_{uA}$ of the *u-th* modality and auxiliary modality feature is obtained by the sigmoid activation function $\sigma(\cdot)$, which is defined as follows.

$$\boldsymbol{S}_{uA} = \sigma(Ali_f(Concat(\boldsymbol{f}^u, \boldsymbol{f}^A))), u \in \{P, S\}, \quad (2)$$

where $Ali_f(\cdot)$ denotes the bilinear align module; $Concat(\cdot)$ is the concatenate operator; u denotes the photo modality or sketch modality.

Based on Eq.(2), the aligned photo feature $\boldsymbol{f}_{ali}^P$ and aligned sketch feature $\boldsymbol{f}_{ali}^S$ can be obtained by Eq.(3).

$$\boldsymbol{f}_{ali}^u = \boldsymbol{S}_{uA} \odot \boldsymbol{f}^u + \boldsymbol{f}^u, u \in \{P, S\}, \quad (3)$$

where $\boldsymbol{f}_{ali}^u$ denotes the aligned feature of the *u-th* modality.

At the same time, the aligned auxiliary feature $\boldsymbol{f}_{ali}^A$ can also be obtained as follows.

$$\boldsymbol{f}_{ali}^A = \boldsymbol{S}_{SA} \odot \boldsymbol{f}^A + \boldsymbol{f}^A. \quad (4)$$

Next, an effective auxiliary cross-attention module (ACA) is presented to achieve information interaction among three modalities, which utilizes auxiliary modality to guide the model to learn the distribution of modality-shared representations. Our ACA can achieve significant information exchange and fusion between photo and sketch modalities.

Taking from photo to auxiliary modality matching as an example, $\boldsymbol{f}_{ali}^P$ and $\boldsymbol{f}_{ali}^A$ are considered as query $\boldsymbol{Q}^P$ and key $\boldsymbol{K}^A$, respectively. Then, the matching weight $\boldsymbol{W}_{P\to A}$ from photo to auxiliary modality can be defined by:

$$\boldsymbol{W}_{P\to A} = Softmax\Big(\frac{\boldsymbol{Q}^P (\boldsymbol{K}^A)^T}{\sqrt{d_K}}\Big), \quad (5)$$

where d_K is the channel dimension of $\boldsymbol{K}^A$;

Similarly, the matching weight $\boldsymbol{W}_{A\to P}$ from auxiliary to photo is also obtained by exchanging query and key.

$$\boldsymbol{W}_{A\to P} = Softmax\Big(\frac{\boldsymbol{Q}^A (\boldsymbol{K}^P)^T}{\sqrt{d_K}}\Big), \quad (6)$$

Based on the above analysis, the photo feature refined by auxiliary modality feature can be defined as follows.

$$\hat{\boldsymbol{f}}^P = Norm(\boldsymbol{W}_{P\to A} \boldsymbol{W}_{A\to P} \boldsymbol{V}^P), \boldsymbol{V}^P = \boldsymbol{f}_{ali}^P, \quad (7)$$

where $Norm(\cdot)$ denotes the layer normalization; $\hat{\boldsymbol{f}}^P$ denotes the photo feature highlighted by the auxiliary modality.

In the same way, we can also obtain the refined sketch feature $\hat{\boldsymbol{f}}^S$ that learns latent representation in auxiliary modality.

Multi-Modality Collaborative Learning

To enhance the robustness of the learned feature relations, we propose a sketch-auxiliary-photo collaborative learning optimization strategy, which consists of the identity classification loss $\mathcal{L}_{ID}$, cross-modality circle loss $\mathcal{L}_{CM}$ and intra-modality circle loss $\mathcal{L}_{IM}$. The proposed learning strategy exploits the auxiliary modality to learn the distribution of sketch and photo modalities at the identity level. First, an identity classification loss is designed to learn the same identity patterns among three modalities, which is defined by:

$$\mathcal{L}_{ID} = \mathcal{L}_{id}(\boldsymbol{F}) + \mathcal{L}_{id}(\hat{\boldsymbol{F}}), \quad (8)$$

where $F = \{\boldsymbol{f}^P, \boldsymbol{f}^A, \boldsymbol{f}^S\}$ denotes the feature map from the output of the shared ResBlocks; $\hat{F} = \{\hat{\boldsymbol{f}}^P, \hat{\boldsymbol{f}}^S\}$ is the feature map from the output of the modality interactive attention module; $\mathcal{L}_{id}$ stands for the cross-entropy classification loss.

Further, we propose an effective cross-modality circle loss to optimize the distance among sketch, photo and auxiliary modalities, which is an improved version of the original circle loss (Sun et al. 2020). The original circle loss can't be effectively utilized to constrain the relationships of different modalities when there is a certain correlation across modalities, which is written as:

$$\begin{aligned}\mathcal{L}_{Cir} &= \mathcal{L}_C(\mathbf{Z}_j^-, \mathbf{Z}_i^+, \gamma, m) \\ &= \log\Big[1 + \sum_{j=1}^{L} \exp(\gamma \alpha_j^- (\mathbf{Z}_j^- - \delta(-))) \\ &\cdot \sum_{i=1}^{H} \exp(-\gamma \alpha_i^+ (\mathbf{Z}_i^+ - \delta(+)))\Big],\end{aligned} \tag{9}$$

where $\delta(-) = m$; $\delta(+) = 1 - m$; $\alpha_j^- = [\mathbf{Z}_j^- + m]_+$; $\alpha_i^+ = [1 + m - \mathbf{Z}_i^+]_+$; $[\cdot]_+$ is the "cut-off at zero" operation to ensure that α_j^- and α_i^+ are non-negative; $\mathbf{Z}_j^+$ and $\mathbf{Z}_i^-$ are the positive and negative sample pairs, respectively; L and H are the numbers of input positive and negative pairs; γ is a scale factor; m denotes the margin parameter.

Based on Eq.(9), we extend it to handle the sketch-photo recognition task, which is described as:

$$\begin{aligned}&\mathcal{L}_{CM}^{uv} = \mathcal{L}_C(\mathbf{Z}_j^{uv-}, \mathbf{Z}_i^{uv+}, \gamma, m_{cm}),\ u \neq v, \\ &\mathbf{Z}_j^{uv-}(\boldsymbol{f}^u, \boldsymbol{f}^v) = \frac{(\boldsymbol{f}_j^u)^T \cdot \boldsymbol{f}^v}{||\boldsymbol{f}_j^u|| \times ||\boldsymbol{f}^v||}, u, v \in \{S, P, A\}, \\ &\mathbf{Z}_i^{uv+}(\boldsymbol{f}^u, \boldsymbol{f}^v) = \frac{(\boldsymbol{f}_i^u)^T \cdot \boldsymbol{f}^v}{||\boldsymbol{f}_i^u|| \times ||\boldsymbol{f}^v||},\end{aligned} \tag{10}$$

where u and v denote any two different modalities in sketch, photo and auxiliary modalities; the features from photo and sketch modalities are $\hat{\boldsymbol{f}}^P$ and $\hat{\boldsymbol{f}}^S$; $\mathbf{Z}_j^{uv-}$ and $\mathbf{Z}_i^{uv+}$ are the cosine similarity of cross-modality negative and positive samples, respectively; m_{cm} stands for the margin parameter to balance the cross-modality distance. Finally, the proposed cross-modality circle loss $\mathcal{L}_{CM}$ is written as:

$$\mathcal{L}_{CM} = \mathcal{L}_{CM}^{AS} + \mathcal{L}_{CM}^{AP} + \mathcal{L}_{CM}^{PS}, \tag{11}$$

Nevertheless, the cross-modality discrepancy is significantly larger than the intra-modality one in cross-modality learning. As a result, it would impel the network swing to optimize cross-modality discrepancy but ignore the intra-modality one. Specifically, $\mathcal{L}_{CM}$ usually pushes visually similar images closer, resulting in sub-optimal latent space. To solve this issue, we further design the intra-modality circle loss $\mathcal{L}_{IM}$ to bring learned global features and modality-shared features of the same identify closer from others within each modality, which is formulated by:

$$\begin{aligned}\mathcal{L}_{IM} &= \mathcal{L}_C(\mathbf{Z}_j^{S-}(\hat{\boldsymbol{f}}^S, \boldsymbol{f}^S), \mathbf{Z}_i^{S+}(\hat{\boldsymbol{f}}^S, \boldsymbol{f}^S), \gamma, m_{im}) \\ &+ \mathcal{L}_C(\mathbf{Z}_j^{A-}(\boldsymbol{f}^A, \boldsymbol{f}^A), \mathbf{Z}_i^{A+}(\boldsymbol{f}^A, \boldsymbol{f}^A), \gamma, m_{im}) \\ &+ \mathcal{L}_C(\mathbf{Z}_j^{P-}(\hat{\boldsymbol{f}}^P, \boldsymbol{f}^P), \mathbf{Z}_i^{P+}(\hat{\boldsymbol{f}}^P, \boldsymbol{f}^P), \gamma, m_{im}),\end{aligned} \tag{12}$$

where $\hat{\boldsymbol{f}}^S$ and $\boldsymbol{f}^S$ are sketch features at different stages in $\mathbf{Z}_i^{S+}(\cdot,\cdot)$, which are the most dissimilar pair with the same identity; $\hat{\boldsymbol{f}}^S$ and $\boldsymbol{f}^S$ are the most similar pair with different identities in $\mathbf{Z}_j^{S-}(\cdot,\cdot)$; $\hat{\boldsymbol{f}}^P$ and $\boldsymbol{f}^P$ share the same principle as $\hat{\boldsymbol{f}}^S$ and $\boldsymbol{f}^S$; m_{im} stands for the relaxation margin to balance the intra-modality distance. In this way, we fully utilize sample patterns to enhance the constraint on intra-modality feature distribution in the case of limited samples, improving the discriminative ability of our model.

The overall objective function $\mathcal{L}$ of our method can be summarized as:

$$\mathcal{L} = \mathcal{L}_{ID} + \mathcal{L}_{CM} + \mathcal{L}_{IM} + \lambda \mathcal{L}_{SR}, \tag{13}$$

where λ is used to control the contribution of $\mathcal{L}_{SR}$ term.

Experiments

Experimental Settings

Datasets. The experiments are performed on five public sketch-based datasets: PKU-Sketch (Pang et al. 2018), QMUL-ShoeV2(ShoeV2) (Yu et al. 2021), QMUL-ChairV2 (ChairV2) (Yu et al. 2021), CUHK student (Tang and Wang 2002) and CUFSF (Zhang, Wang, and Tang 2011).

PKU-Sketch is the first public Sketch Re-ID dataset, which consists of 200 pedestrians and each identity has two photos and one sketch. According to the experimental setting in (Pang et al. 2018), 150 identities are utilized for training and the rest 50 identities for testing. ShoeV2 and ChairV2 datasets are commonly exploited for sketch-based image retrieval task. ShoeV2 consists of 2000 photos and 6730 sketches with 2000 identities in total. ChairV2 includes 400 photos and 1275 sketches, with one ID for each photo. The details of the training and testing sets can be found in (Yu et al. 2021). In addition, we further prove the generalization performance of the proposed method on the CUHK student and CUFSF datasets. The CUHK student dataset has 188 face photo-sketch pairs, of which 88 pairs are selected for training and the rest 100 pairs for testing. For CUFSF dataset, consisting of 1194 face photo-sketch pairs, we randomly select 500 and 694 persons as train and test sets according to (Zhang, Wang, and Tang 2011), respectively.

Evaluation Metrics. Following the existing settings in (Pang et al. 2018; Zhang et al. 2022), we evaluate the proposed methods by standard Cumulative Matching Characteristic (CMC) at Rank-k and mean Average Precision (mAP).

Implementation Details. We implement our model on Pytorch framework with an NVIDIA RTX-3090 GPU. In this paper, the ResNet-50 (He et al. 2016) pre-trained by ImageNet is adopted as our backbone. The first two stages of ResNet-50 are adopted as the modality-specific ResBlocks, and the rest of the stages are used as the shared ResBlocks. During the training stage, data augmentations, including random horizontal flipping, padding and random cropping, are used to prevent overfitting. We employ the AdamW optimizer for 100 epochs with an initial learning rate of 0.00045 and a weight decay of 0.02. In the 3×3 convolutional kernel of our generator, the initial center weight is set to 9 and the surrounding weights are set to -0.8. The parameters ξ and λ are set to 0.07 and 0.6, respectively. The scale parameter γ

B	Aux.	$\mathcal{L}_{Cir}$	$\mathcal{L}_{CM}$	MIA	$\mathcal{L}_{IM}$	PKU-Sketch R1	PKU-Sketch mAP	ShoeV2 R1	ShoeV2 mAP
✓						67.8	68.0	29.1	40.8
✓	✓	✓				71.6	72.9	32.7	46.1
✓	✓		✓			76.5	77.3	37.5	50.3
✓	✓		✓	✓		86.4	84.9	45.8	58.2
✓	✓		✓	✓	✓	90.0	86.2	48.5	60.6

Table 1: Evaluation of each component of the proposed method on PKU-Sketch and ShoeV2 datasets. CMC (%) at Rank1 (R1) and mAP (%).

is 64. The margin parameters m_{cm} and m_{im} are set to 0.25 and 0.5, respectively.

Ablation Study

Effectiveness of Each Component. We evaluate the performance of each component on PKU-Sketch and ShoeV2 datasets in Table 1. In this experiment, we use the ResNet-50 trained with identity loss as Baseline, denoted as B. Compared with the Baseline, our auxiliary modality (Aux.) can dynamically relieve cross-modality discrepancy with cross-modality circle loss ($\mathcal{L}_{CM}$), thus greatly improving all the metrics. Additionally, we evaluate the original circle loss ($\mathcal{L}_{Cir}$) to verify the effectiveness of the proposed $\mathcal{L}_{CM}$. Further, the performance is improved by introducing the modality interactive attention module (MIA) to refine the photo and sketch modalities. Meanwhile, the proposed intra-modality circle loss ($\mathcal{L}_{IM}$) helps discriminative learning. By combining all components, we can regulate a more robust cross-modality feature relationship, achieving a rank-1 of 90.0% and mAP of 86.2% on the PKU-Sketch dataset. The results demonstrate that all the proposed components contribute consistently to the accuracy gain.

Effectiveness of Dynamic Auxiliary Generator. The dynamic auxiliary generator (DAG) plays a crucial role in generating auxiliary modality to bridge the gap between sketch modality and photo modality. The existing method (Chen et al. 2022) utilizes a ready-made generator (Xiang et al. 2022) to generate sketch-like images as an auxiliary modality, which is not well correlated with the recognition task and produces much sketch-like background noise. Thus, we introduce our generator into the sketch recognition task for joint training (G_j) with style refinement loss ($\mathcal{L}_{SR}$). To demonstrate the effectiveness of our method, we isolate our generator and fix its parameters without $\mathcal{L}_{SR}$ (G_f). As shown in Table 2, DAG under the constraint of $\mathcal{L}_{SR}$ ($G_j + \mathcal{L}_{SR}$) obtains a significant improvement with a Rank-1 accuracy of 90.0% on PKU-Sketch dataset.

Comparison with State-of-the-Art Methods

In this subsection, we compare our DALNet with the existing methods on the PKU-Sketch dataset, and the results are shown in Table 3. We can see that our DALNet achieves the Rank-1 accuracy of 90.0% and mAP of 86.2%. Compared to traditional adversarial learning (Pang et al. 2018) and feature fusion (Gui et al. 2020; Zhu et al. 2022) methods, DALNet outperforms them by a margin of at least 29.2% Rank-1. The reason can be attributed that our method can alleviate both the cross-modality difference and intra-modality variation. Additionally, DALNet is also superior to existing synthesis learning methods (Chen et al. 2022) thanks to the robust auxiliary modality, which is well correlated with the recognition task and has a better quality without extra noise. Furthermore, DALNet outperforms CCSC (Zhang et al. 2022) by 4% Rank-1 and 2.5% mAP. Based on the above analysis, our DALNet can learn more invariant latent representations and align the feature distribution of three modalities.

G_f	G_j	$\mathcal{L}_{SR}$	PKU-Sketch R1	PKU-Sketch R5	PKU-Sketch mAP	ShoeV2 R1	ShoeV2 R5	ShoeV2 mAP
✓			86.5	98.4	84.4	45.2	72.9	57.9
	✓		88.4	98.2	85.6	46.8	75.6	59.9
	✓	✓	90.0	98.6	86.2	48.5	76.4	60.6

Table 2: Performance comparison of fixing (G_f) or jointly optimizing generator on PKU-Sketch dataset. G_j refers to the joint training of generation task and recognition task. CMC (%) at Rank-1 (R1), Rank-5 (R5) and mAP (%).

Methods	R1	R5	R10	mAP
TripletSN (Yu et al. 2016)	9.0	26.8	42.2	-
GNSiamese (Sangkloy et al. 2016)	28.9	54.0	62.4	-
AFLNet (Pang et al. 2018)	34.0	56.3	72.5	-
LMDI (Gui et al. 2020)	49.0	70.4	80.2	-
CDAC (Zhu et al. 2022)	60.8	80.6	88.8	-
SketchTrans (Chen et al. 2022)	84.6	94.8	98.2	-
CCSC (Zhang et al. 2022)	86.0	98.0	100.0	83.7
DALNet (Ours)	90.0	98.6	100.0	86.2

Table 3: Comparison with state-of-the-art methods in terms of CMC (%) and mAP (%) on PKU-Sketch dataset.

Evaluation on Generalizability

To verify the generalization ability of the proposed DALNet, we conduct experiments on the sketch-based image retrieval and sketch-photo face recognition tasks.

Sketch-Based Image Retrieval. As shown in Table 4, the proposed method surpasses existing SOTAs on ShoeV2 and ChairV2 datasets in the sketch-based image retrieval task. Specifically, our DALNet reaches the Rank-1 accuracy of 82.3% and Rank-10 of 98.3% on the ChairV2 dataset, improving the Rank-1 by 0.6% and Rank-10 by 0.9% over the sub-optimal methods. Moreover, we achieve Rank-1 of 48.5% and Rank-10 of 86.1% on the ShoeV2 dataset.

Sketch-Photo Face Recognition. The evaluations are listed in Table 5, the best performance on CUHK student and CUFSF datasets also demonstrate the excellent generalization of our method. For a fair comparison, the evaluation metric used is Rank-1, the same as SOTAs. Our DALNet obtains a Rank-1 of 96.50% on CUHK student dataset, which outperforms the state-of-the-art CDAC by 2.5%. In addition, we achieve Rank-1 of 97.84% on the CUFSF dataset.

Methods	ChairV2		ShoeV2	
	R1	R10	R1	R10
TripletSN (Yu et al. 2016)	47.4	84.3	28.7	71.6
HOLEF SN (Song et al. 2017a)	50.7	86.3	31.2	74.6
SN-RL (Collomosse et al. 2017)	51.2	86.9	30.8	74.2
DVML (Lin et al. 2018)	52.8	85.2	32.1	76.2
SSL (Bhunia et al. 2021)	53.3	87.5	33.4	80.7
Triplet attn (Song et al. 2017b)	53.4	87.6	31.7	75.8
CC-Gen (Pang et al. 2019)	54.2	88.2	33.8	77.9
TripletRL (Bhunia et al. 2020)	56.5	89.6	34.1	78.8
Stylemeup (Sain et al. 2021)	62.9	91.1	36.5	81.8
RLF (Bhunia et al. 2022)	64.8	-	43.7	-
CCSC (Zhang et al. 2022)	74.3	97.4	33.5	80.2
Sketch-PVT (Sain et al. 2023)	74.6	92.7	48.3	85.6
SketchTrans (Chen et al. 2022)	81.7	97.4	38.7	80.9
DALNet (Ours)	82.3	98.3	48.5	86.1

Table 4: Comparison with state-of-the-art methods in terms of CMC (%) on ChairV2 and ShoeV2 datasets.

Methods	CUHK	CUFSF
	R1	R1
DR-GAN (Tran, Yin, and Liu 2017)	83.70	-
Dual-Transfer (Zhang et al. 2018)	86.30	-
KD (Zhu et al. 2020)	93.55	66.37
IA CycleGAN (Fang et al. 2020)	93.62	64.94
G-HFR (Peng et al. 2016)	-	96.00
MMTN (Luo et al. 2022)	-	97.20
CDAC (Zhu et al. 2022)	94.00	96.80
DALNet (Ours)	96.50	97.84

Table 5: Comparison with state-of-the-art methods in terms of Rank-1 (%) on CUHK student and CUFSF datasets.

Visualization

Visualization of Attention Feature Map. In order to validate the effectiveness of our method, we conduct two sketch queries from different viewpoints to compare the attention feature maps of Baseline and our DALNet by XGrad-CAM (Fu et al. 2020). The top-4 photo ranking results are shown in Figure 3. In the first row, the attention maps obtained by Baseline overlook the relevance of two modalities, especially when the scene and pose change. Moreover, some similar features cause disturbance when the model focuses on the same local regions. In the second row, we can see that our DALNet can focus on multiple similar areas in the sketches and photos, such as key facial features, clothing patterns, bags and chest tags, etc. In general, DALNet can enhance cross-modality retrieval performance by focusing on multiple effective modality-shared attention regions.

Visualization of Feature Distribution. To further analyze the effectiveness of DALNet, we randomly select 10 identities from PKU-Sketch dataset and visualize their 3D feature distributions during the training, as shown in Figure 4. At the initial stage (epoch 0), there are substantial differences between photos (orange dots) and sketches (grey dots), hindering cross-modality matching. As the training progresses, the proposed auxiliary modality (green dots) acts as a bridge to connect the photo and sketch modalities in the common feature space. In Figure 4(b), we can observe that the photo and sketch features gradually converge. Further, the network learns the invariant patterns by regulating smaller intra-class distances and larger inter-class distances at epoch 50 in Figure 4(c). Finally, the auxiliary features are grouped into their respective identity centers in Figure 4(d), demonstrating powerful discriminability under cross-modality scenes. This proves that our DALNet can effectively learn identity-aware features and match the distribution of each identity.

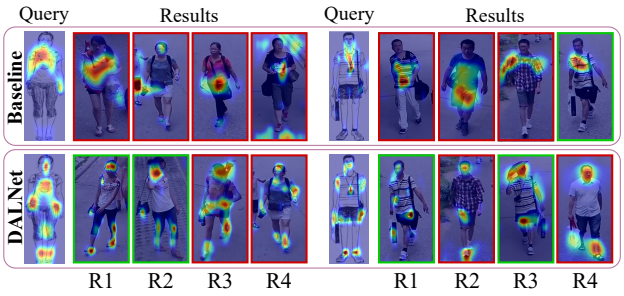
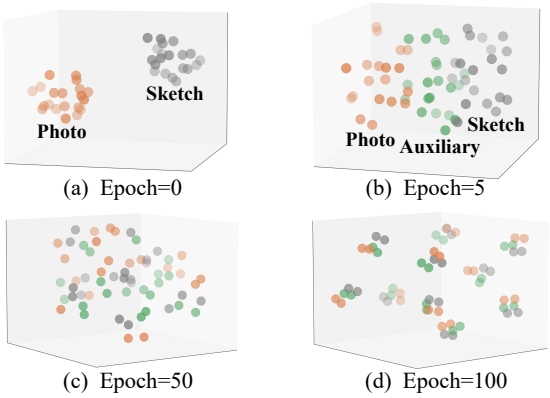


Figure 3: Attention visualization of Baseline and our DALNet on PKU-Sketch dataset. Attention maps from sketch query images and four photo retrieval results are visualized in different views. The red and green rectangles indicate fault retrieval and correct retrieval, respectively.



(a) Epoch=0 (b) Epoch=5

(c) Epoch=50 (d) Epoch=100

Figure 4: Distributions visualization of photo, sketch and auxiliary features by T-SNE (Van der Maaten and Hinton 2008). Different brightness indicates different identities.

Conclusions

In this paper, we propose a differentiable auxiliary learning network by constructing an auxiliary modality to jointly explore the cross-modality feature relationships for Sketch Re-ID. The auxiliary modality is generated by a dynamic auxiliary generator with style refinement loss, guiding modality interactive attention module to align the features and learn invariant patterns of photo and sketch modalities. In addition, the multi-modality collaborative learning scheme is designed to learn the feature relationships for three modalities at the identity level with limited samples. Extensive experiments on five sketch-based datasets demonstrate the effectiveness and generalizability of the proposed method.

Acknowledgments

This work was supported by the Research Council of Finland (former Academy of Finland) Academy Professor project EmotionAI (Grants 336116, 345122), ICT 2023 project TrustFace (Grant 345948), the University of Oulu & Research Council of Finland Proff 7 (Grant 352788), by Infotech Oulu, and National Natural Science Foundation of China (Grants 61802058, 61911530397).

References

Bhunia, A. K.; Chowdhury, P. N.; Sain, A.; Yang, Y.; Xiang, T.; and Song, Y.-Z. 2021. More photos are all you need: Semi-supervised learning for fine-grained sketch based image retrieval. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, 4247–4256.

Bhunia, A. K.; Koley, S.; Khilji, A. F. U. R.; Sain, A.; Chowdhury, P. N.; Xiang, T.; and Song, Y.-Z. 2022. Sketching without worrying: Noise-tolerant sketch-based image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 999–1008.

Bhunia, A. K.; Yang, Y.; Hospedales, T. M.; Xiang, T.; and Song, Y.-Z. 2020. Sketch less for more: On-the-fly fine-grained sketch-based image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9779–9788.

Bui, T.; Ribeiro, L.; Ponti, M.; and Collomosse, J. 2017. Compact descriptors for sketch-based image retrieval using a triplet loss convolutional neural network. *Computer Vision and Image Understanding*, 164: 27–37.

Chen, C.; Ye, M.; Qi, M.; and Du, B. 2022. Sketch Transformer: Asymmetrical Disentanglement Learning from Dynamic Synthesis. In *Proceedings of the 30th ACM International Conference on Multimedia*, 4012–4020.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A Simple Framework for Contrastive Learning of Visual Representations. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.

Chen, Y.; Zhang, Z.; Wang, Y.; Zhang, Y.; Feng, R.; Zhang, T.; and Fan, W. 2021. AE-Net: Fine-grained sketch-based image retrieval via attention-enhanced network. *Pattern Recognition*, 108291.

Collomosse, J.; Bui, T.; and Jin, H. 2019. Livesketch: Query perturbations for guided sketch-based visual search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2879–2887.

Collomosse, J.; Bui, T.; Wilber, M. J.; Fang, C.; and Jin, H. 2017. Sketching with style: Visual search with sketches and aesthetic context. In *Proceedings of the IEEE international conference on computer vision*, 2660–2668.

Deng, C.; Xu, X.; Wang, H.; Yang, M.; and Tao, D. 2020. Progressive cross-modal semantic network for zero-shot sketch-based image retrieval. *IEEE Transactions on Image Processing*, 29: 8892–8902.

Fang, Y.; Deng, W.; Du, J.; and Hu, J. 2020. Identity-aware CycleGAN for face photo-sketch synthesis and recognition. *Pattern Recognition*, 102: 107249.

Fu, R.; Hu, Q.; Dong, X.; Guo, Y.; Gao, Y.; and Li, B. 2020. Axiom-based grad-cam: Towards accurate visualization and explanation of cnns. *arXiv preprint arXiv:2008.02312*.

Gui, S.; Zhu, Y.; Qin, X.; and Ling, X. 2020. Learning multi-level domain invariant features for sketch re-identification. *Neurocomputing*, 403: 294–303.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Li, Y.; Hospedales, T.; Song, Y.-Z.; and Gong, S. 2014. Fine-grained sketch-based image retrieval by matching deformable part models. *British Machine Vision Conference*.

Lin, X.; Duan, Y.; Dong, Q.; Lu, J.; and Zhou, J. 2018. Deep variational metric learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 689–704.

Luo, M.; Wu, H.; Huang, H.; He, W.; and He, R. 2022. Memory-modulated transformer network for heterogeneous face recognition. *IEEE Transactions on Information Forensics and Security*, 17: 2095–2109.

Pang, K.; Li, K.; Yang, Y.; Zhang, H.; Hospedales, T. M.; Xiang, T.; and Song, Y.-Z. 2019. Generalising fine-grained sketch-based image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 677–686.

Pang, K.; Song, Y.-Z.; Xiang, T.; and Hospedales, T. M. 2017. Cross-domain Generative Learning for Fine-Grained Sketch-Based Image Retrieval. In *BMVC*, 1–12.

Pang, K.; Yang, Y.; Hospedales, T. M.; Xiang, T.; and Song, Y.-Z. 2020. Solving Mixed-Modal Jigsaw Puzzle for Fine-Grained Sketch-Based Image Retrieval. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Pang, L.; Wang, Y.; Song, Y.-Z.; Huang, T.; and Tian, Y. 2018. Cross-domain adversarial feature learning for sketch re-identification. In *Proceedings of the 26th ACM international conference on Multimedia*, 609–617.

Peng, C.; Gao, X.; Wang, N.; and Li, J. 2016. Graphical representation for heterogeneous face recognition. *IEEE transactions on pattern analysis and machine intelligence*, 39(2): 301–312.

Qin, X.; Zhang, Z.; Huang, C.; Dehghan, M.; Zaiane, O. R.; and Jagersand, M. 2020. U2-Net: Going deeper with nested U-structure for salient object detection. *Pattern recognition*, 106: 107404.

Radenović, F.; Tolias, G.; and Chum, O. 2018. Fine-tuning CNN image retrieval with no human annotation. *IEEE transactions on pattern analysis and machine intelligence*, 41(7): 1655–1668.

Sain, A.; Bhunia, A.; Yang, Y.; Xiang, T.; and Song, Y.-Z. 2020. Cross-Modal Hierarchical Modelling for Fine-Grained Sketch Based Image Retrieval. *British Machine Vision Conference*.

Sain, A.; Bhunia, A. K.; Koley, S.; Chowdhury, P. N.; Chattopadhyay, S.; Xiang, T.; and Song, Y.-Z. 2023. Exploiting

Unlabelled Photos for Stronger Fine-Grained SBIR. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6873–6883.

Sain, A.; Bhunia, A. K.; Yang, Y.; Xiang, T.; and Song, Y.-Z. 2021. Stylemeup: Towards style-agnostic sketch-based image retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8504–8513.

Sangkloy, P.; Burnell, N.; Ham, C.; and Hays, J. 2016. The sketchy database: learning to retrieve badly drawn bunnies. *ACM Transactions on Graphics (TOG)*, 35(4): 1–12.

Song, J.; Pang, K.; Song, Y.-Z.; Xiang, T.; and Hospedales, T. M. 2018. Learning to Sketch with Shortcut Cycle Consistency. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Song, J.; Song, Y.-Z.; Xiang, T.; and Hospedales, T. M. 2017a. Fine-Grained Image Retrieval: the Text/Sketch Input Dilemma. In *BMVC*, volume 2, 7.

Song, J.; Yu, Q.; Song, Y.-Z.; Xiang, T.; and Hospedales, T. M. 2017b. Deep spatial-semantic attention for fine-grained sketch-based image retrieval. In *Proceedings of the IEEE international conference on computer vision*, 5551–5560.

Sun, Y.; Cheng, C.; Zhang, Y.; Zhang, C.; Zheng, L.; Wang, Z.; and Wei, Y. 2020. Circle loss: A unified perspective of pair similarity optimization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6398–6407.

Tang, X.; and Wang, X. 2002. Face photo recognition using sketch. In *Proceedings. International Conference on Image Processing*, volume 1, I–I. IEEE.

Tran, L.; Yin, X.; and Liu, X. 2017. Disentangled representation learning gan for pose-invariant face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1415–1424.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

Xiang, X.; Liu, D.; Yang, X.; Zhu, Y.; Shen, X.; and Allebach, J. P. 2022. Adversarial open domain adaptation for sketch-to-photo synthesis. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 1434–1444.

Ye, M.; Shen, J.; Lin, G.; Xiang, T.; Shao, L.; and Hoi, S. C. 2021. Deep learning for person re-identification: A survey and outlook. *IEEE transactions on pattern analysis and machine intelligence*, 44(6): 2872–2893.

Yu, Q.; Liu, F.; Song, Y.-Z.; Xiang, T.; Hospedales, T. M.; and Loy, C. C. 2016. Sketch Me That Shoe. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yu, Q.; Song, J.; Song, Y.-Z.; Xiang, T.; and Hospedales, T. M. 2021. Fine-grained instance-level sketch-based image retrieval. *International Journal of Computer Vision*, 129: 484–500.

Zhang, M.; Wang, R.; Gao, X.; Li, J.; and Tao, D. 2018. Dual-transfer face sketch–photo synthesis. *IEEE Transactions on Image Processing*, 28(2): 642–657.

Zhang, W.; Wang, X.; and Tang, X. 2011. Coupled information-theoretic encoding for face photo-sketch recognition. In *CVPR 2011*, 513–520. IEEE.

Zhang, Y.; Wang, Y.; Li, H.; and Li, S. 2022. Cross-Compatible Embedding and Semantic Consistent Feature Construction for Sketch Re-identification. In *Proceedings of the 30th ACM International Conference on Multimedia*, 3347–3355.

Zhu, F.; Zhu, Y.; Jiang, X.; and Ye, J. 2022. Cross-Domain Attention and Center Loss for Sketch Re-Identification. *IEEE Transactions on Information Forensics and Security*, 17: 3421–3432.

Zhu, M.; Li, J.; Wang, N.; and Gao, X. 2020. Knowledge distillation for face photo–sketch synthesis. *IEEE Transactions on Neural Networks and Learning Systems*, 33(2): 893–906.