2025 International Conference on Intelligent Computing
July 26-29, Ningbo, China
https://www.ic-icc.cn/2025/index.php

Cross-Modal Dependable Subjective Learning for Sketch Person Re-identification

Junjie Huang[1], Chuang Li[1], Zhihong Sun[2*]

[1] Wuhan Textile University, Wuhan, 430200, China.
[2] Naval University of Engineering, Wuhan, 430033, China.

Abstract. Sketch-based person re-identification (Sketch Re-ID) enables suspect retrieval when camera images are unavailable by leveraging sketches drawn from human memory. However, the subjectivity in sketches often introduces significant style variation, making it difficult to extract dependable cross-modal features. To address this challenge, we propose a novel Cross-Modal Dependable Subjective Learning (CMDSL) framework. It consists of a Flexible Feature Aggregation Module (FFAM) that removes style noise via instance normalization and captures dependable subjective semantics through attention-enhanced residual learning, and a Recognisable Target Centroid Loss (RTCL) that strengthens discriminability and alignment across modalities. Experiments on MARKET-SKETCH-1K and PKU-Sketch datasets demonstrate that our approach effectively captures consistent subjective cues and achieves state-of-the-art performance under diverse sketch styles.

Keywords: Person re-identification, Sketch retrieval, Dependable Subjective Features, Target Centroid Loss.

1 Introduction

Person re-identification (Re-ID) has important prospects in many real-world applications that focus on identifying individuals from surveillance data, particularly through image analysis techniques. Traditional methods [1] for Person Re-ID primarily rely heavily on RGB images captured from surveillance cameras to identify individuals based on their appearance. However, they are not applicable when only eyewitness descriptions [2] and hand-drawn sketches [3] of individuals are available. It will lead to the emerging field of Sketch-based person Re-ID, which matches hand-drawn sketches with gallery images to identify individuals.

In recent years, sketch-based Re-ID has gained increasing attention due to its wide range of applications in video surveillance, law enforcement, and human-computer interaction. This mode transition brings unique challenges: 1) domain gaps between sketches and images. 2) significant intra-class variation and limited discriminative information in sketches. According to the report of the Los Angeles store robbery [4], eyewitnesses' descriptions of the suspect's appearance vary and are often incomplete, which may be influenced by subjective bias. Different perspectives and perceptual nuances among witnesses may lead to conflicting descriptions, complicating the task of

*Corresponding author is Zhihong Sun (`zhihong.sun@whu.edu.cn`) accurately portraying a suspect. In addition, artists' sketches may be biased by individual aesthetic tendencies and artistic styles. Therefore, the 3th challenge is to deal with and understand these overlooked subjective factors.

To address the aforementioned limitations, we propose a novel framework termed Cross-Modal Dependable Subjective Feature Learning (CMDSL) in this paper. The proposed framework aims to extract dependable subjective information from modality-specific features and refine them into modality-shared representations, thereby enhancing their discriminative capacity for cross-modal matching tasks. Specifically, we employ a dual-stream network [5] architecture to independently extract modality-shared features from each modality. However, it is important to note that the shared features obtained via dual-stream networks often suffer from modality inconsistency and lack of reliability. To alleviate this, we introduce a Flexible Feature Aggregation Module (FFAM), which performs style removal via instance normalization, structural detail recovery through a residual branch, and dependable information enhancement through global context modeling and channel attention mechanisms. This enables the model to precisely capture those structural semantic cues that are consistent and subjectively reliable across modalities. Furthermore, we design a Recognisable Target Centroid Loss (RTCL) to further improve the intra-class compactness and inter-class separability of the extracted subjective features in the embedding space.

Overall, the contributions of this work can be summarized as follows:

- We propose the Cross-Modal Dependable Subjective Feature Learning (CMDSL) network, which utilizes the implicit subjective information in specific modal features to enhance retrieval capabilities for the sketch Re-ID task.
- The proposed FFAM extracts dependable, modality-invariant subjective features by disentangling style and structure information and enhancing discriminative semantics via attention. While the RTCL further improves feature compactness and class separability by guiding features toward consistent centroid alignment across modalities.
- Finally, extensive experiments on the MARKET-SKETCH-1K [6] and PKU Sketch Re-ID [3] datasets demonstrate the effectiveness and robustness of the proposed method compared to state-of-the-art approaches.

2 Related Work

2.1 Cross-modal image retrieval

Cross-modal image retrieval seeks to match related content across different data types, such as sketches, RGB photos, infrared images, and text. Early re-identification methods [7] often use a Siamese network to learn modality-specific embeddings and pull matching pairs together. Some works [8] fuse features from separate branches to enrich representations, but simple fusion can cause information loss or confusion. More recent approaches employ large vision–language models and transformer-based alignment. LG-MGC [9] enhances CLIP embeddings with local semantic completion and generative translation to recover fine details and shrink the modality gap. HAT [10] uses identical transformer encoders for both image and text, aligning multi-level features to produce more compatible embeddings. These methods move beyond basic Siamese or fusion schemes toward richer, attention-driven alignment for robust cross-modal matching.

2.2 Sketch Re-Identification

Sketch-based person re-identification (Sketch Re-ID) is a challenging task that aims to match hand-drawn sketches with corresponding photographs. Unlike conventional image retrieval tasks, Sketch Re-ID demands strong generalization to unseen sketches, making standard approaches less effective. This difficulty primarily stems from the substantial modality gap between sketches and photos—especially in terms of detail and style—which becomes even more pronounced when the sketches are highly abstract or simplified. To address this, recent studies [11,12] have proposed various strategies to bridge the sketch-photo domain gap. However, traditional methods often struggle to capture consistent semantic information across modalities. While some approaches [13] leverage generative metric learning to enhance robustness, they can introduce high computational overhead and may lead to unreliable or noisy semantic features.

3 The proposed method

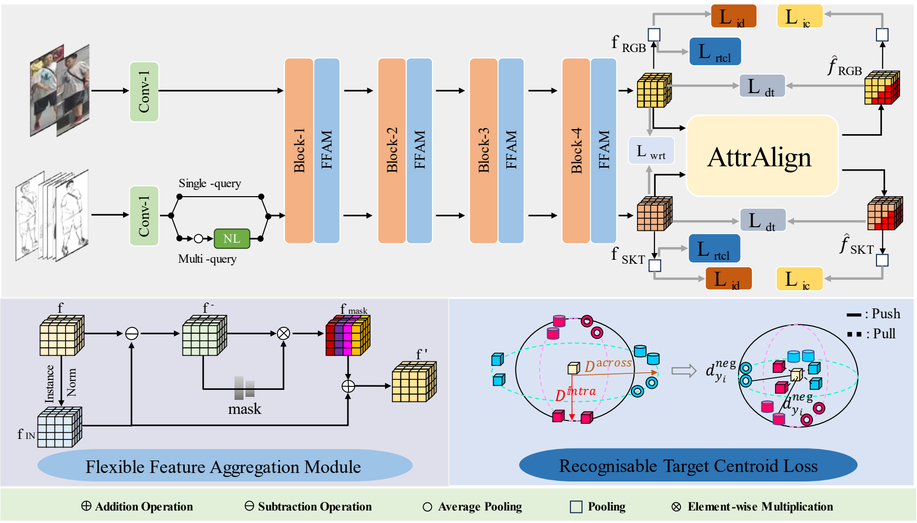

Fig. 1. Framework of the proposed Cross-Modal Dependable Subjective Learning model. The Framework includes a Flexible Feature Aggregation Module and a Recognisable Target Centroid Loss.

In this section, we elaborate on the proposed Cross-Modal Dependable Subjective Feature Learning (CMDSL) framework, as illustrated in Fig.1. CMDSL is designed to address the intrinsic subjectivity and stylistic discrepancies between sketches and visible images in the sketch-based person re-identification task. Specifically, CMDSL adopts a dual-stream architecture, where each stream independently extracts features from its respective modality—sketches and RGB images. To mitigate the unreliability of raw modality-shared features extracted from different domains, we integrate the Flexible Feature Aggregation Module (FFAM) after each ResNet block. FFAM performs instance normalization to strip away stylistic variations while preserving discriminative identity cues via residual reconstruction and global channel-wise attention, enabling the model to selectively emphasize dependable subjective information that remains consistent across modalities. In addition, we introduce the Recognisable Target Centroid Loss (RTCL) to further guide the model in focusing on consistent identity representations. RTCL leverages mini-batch feature distributions to penalize unreliable cross-modal deviations and dynamically adjusts the inter-class separation margins, thereby promoting compact intra-class clusters and improving discriminability. Together, these components enable CMDSL to extract robust, subjective yet dependable features that generalize well across artistically diverse sketches and natural images.

Formally, let $\mathcal{R} = \{x_i^R\}_{i=1}^{N_R}$ and $\mathcal{S} = \{x_i^S\}_{i=1}^{N_S}$ represent the sets of RGB and sketch images in the dataset, respectively. In most cases, an equal number of samples are drawn from both modalities for each training mini-batch. That is, $N_R = N_S = N = P \times K$, where P denotes the number of distinct person identities, and K indicates the number of images selected per identity for each modality. Under this setting, each mini-batch contains $2N$ images in total,N from RGB and N from sketch modality. The combined mini-batch image set is represented as $\mathcal{X} = \{x_i \mid x_i \in \mathcal{R} \cup \mathcal{S}\}_{i=1}^{2N}$, and the corresponding label set is denoted by $\mathcal{Y} = \{y_i\}_{i=1}^{P\times K}$.Subsequently, a dual-stream network equipped with the FFAM is employed to independently extract modality-specific features from RGB and sketch images.

3.1 Flexible Feature Aggregation Module

Sketch-based Re-ID must overcome a significant cross-modal gap: sketches are abstract line drawings without color or texture, while photos contain rich visual details. This gap is further widened by the diverse drawing styles of different artists, which can strip sketches of fine-grained, identity-consistent cues (e.g., subtle patterns or accessories) even as coarse attributes (e.g., clothing silhouette, hairstyle) remain. To tackle this, we introduce the FFAM, which disentangles style and structure in intermediate features so as to suppress modality-specific noise and preserve dependable subjective cues.

First, FFAM applies Instance Normalization(IN) [14] to the input feature map $\mathbf{f} \in \mathbb{R}^{C\times H\times W}$, removing per-channel style variations:

$$f_{\mathrm{IN}} = \frac{\mathbf{f} - \mathbb{E}[\mathbf{f}]}{\sqrt{\mathrm{Var}[\mathbf{f}] + \epsilon}}, \tag{1}$$

where $\mathbb{E}[\cdot]$ and $\mathrm{Var}[\cdot]$ denote channel-wise mean and variance, and ϵ prevents division by zero. The normalized feature f_{IN} captures modality-invariant structural content. We then recover the discarded style residual via

$$f^{-} = \mathbf{f} - f_{\mathrm{IN}}. \qquad (2)$$

To selectively reintroduce only identity-relevant style components, FFAM computes a channel attention mask,like SE-Net [15] :

$$m = \sigma(L_2\,\mathrm{ReLU}(L_1\,g(\mathbf{f}))), \qquad (3)$$

where $g(\cdot)$ denotes global pooling, $L_1 \in \mathbb{R}^{\frac{C}{r}\times C}$ and $L_2 \in \mathbb{R}^{C\times\frac{C}{r}}$ are learnable projections with reduction ratio r, and σ is the sigmoid activation function. The resulting $m \in [0,1]^C$ weighs each channel of the style residual:

$$f_{\mathrm{mask}} = m \otimes f^{-}, \qquad (4)$$

with $\otimes$ denoting element-wise multiplication. Finally, the refined representation combines the normalized structure and the masked style:

$$f' = f_{\mathrm{IN}} \oplus f_{\mathrm{mask}}, \qquad (5)$$

where $\oplus$ is element-wise addition. By this decomposition and selective fusion, FFAM yields features that are both structure-centric and adaptively enriched with discriminative stylistic details, effectively narrowing the modality gap without reintroducing noise.

3.2 Recognisable Target Centroid Loss

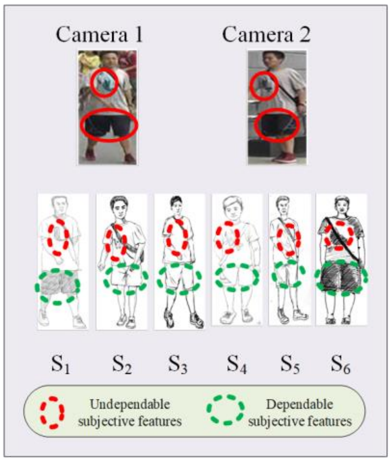
Camera 1
Camera 2
S_1
S_2
S_3
S_4
S_5
S_6
Undependable subjective features
Dependable subjective features

Fig. 2. The motivation for proposing a loss function that captures dependable subjective features across modalities. S1-S6 denote sketches made by different artists for visible light images.

In sketch-based person Re-ID, sketches and photos exhibit substantial appearance gaps, and sketches themselves vary widely in detail depending on the artist's drawing style. As depicted in Fig. 2, while some sketches accurately capture fine-grained identity cues (e.g., T-shirt prints), others omit them, undermining the reliable subjective information extracted by FFAM. To reinforce these dependable subjective features and suppress unstable, sketch-specific artifacts, we introduce the Recognisable Target Centroid Loss

(RTCL). RTCL is designed to 1) penalize excessive deviation between intra-modal and cross-modal feature distances for the same identity, and 2) enforce a dynamic margin based on negative–centroid distances to improve inter-class separability.

Within each mini-batch, we randomly select P identities and then choose K visible light images and K sketch images per identity. Specifically, the anchor features of the sketch modality and visible light modality are defined as f_a^{skt} and f_a^{rgb}, respectively. Taking f_a^{skt} as an example, we compute its average distance to other sketches (intra-modal) and to paired photos (cross-modal):

$$D_a^{\text{intra}} = \frac{1}{K-1}\sum_{\substack{i=1 \\ i\neq a}}^{K} \phi\left(f_a^{\text{skt}}, f_i^{\text{skt}}\right), \tag{6}$$

$$D_a^{\text{across}} = \frac{1}{K}\sum_{i=1}^{K} \phi\left(f_a^{\text{skt}}, f_i^{\text{rgb}}\right), \tag{7}$$

where $\phi(\cdot,\cdot)$ denotes a distance metric (e.g., Euclidean). We then penalize discrepancies between these distances, encouraging stable subjective cues to manifest consistently across modalities:

$$\varphi \;=\; \frac{1}{2PK}\sum_{p=1}^{P}\sum_{a=1}^{2K}\left(D_a^{\text{intra}} - D_a^{\text{across}}\right)^2. \tag{8}$$

To further amplify inter-class separability of these dependable features, we compute each identity's centroid across both modalities:

$$c_{y_i} = \frac{1}{2K}\left(\sum_{j=1}^{K} f_j^{\text{rgb}} + \sum_{k=1}^{K} f_k^{\text{skt}}\right), \tag{9}$$

where c_{y_i}denotes the feature center of the identity y_i, y_i is the identity label of the i-th image.and measure its average distance to all negative examples:

$$d_{y_i}^{\text{neg}} = \frac{1}{2K(P-1)}\sum_{y_j\neq y_i} \| f_j - c_{y_i} \|_2, \tag{10}$$

we then define the adaptive margin term θ as the ratio of (a) the mean positive-centroid distance to (b) the mean of those negative distances below $d_{y_i}^{\text{neg}}$:

$$\theta = \frac{\sum_{i=1}^{P} \mathbb{E}_{y_j=y_i} \| f_j - c_{y_i} \|_2}{\sum_{i=1}^{P} \mathbb{E}_{\substack{y_k\neq y_i \\ \|f_k - c_{y_i}\|_2 < d_{y_i}^{\text{neg}}}} \| f_k - c_{y_i} \|_2}. \tag{11}$$

Finally, RTCL integrates both consistency and margin objectives:

$$\mathcal{L}_{\text{RTCL}} = \varphi \;+\; \theta. \tag{12}$$

thereby reinforcing the dependable subjective features identified by FFAM and promoting robust, discriminative embeddings for sketch-to-photo retrieval.

3.3 Overall Objective Function

The overall objective function used to optimize the network during the training stage is formulated as follows:

$$\mathcal{L} = \mathcal{L}_{id} + \mathcal{L}_{wrt} + \lambda_1(\mathcal{L}_{ic} + \mathcal{L}_{dt}) + \lambda_2\mathcal{L}_{rtcl}, \tag{13}$$

where $\mathcal{L}_{id}$ is the identity (consistency) loss, which plays a crucial role in training as it ensures alignment between the input image and its corresponding label throughout the optimization process. $\mathcal{L}_{wrt}$ is the weighted regularized triplet loss, which optimizes relative distances between positive and negative samples without the need for additional boundary parameters. $\mathcal{L}_{ic}$ promotes consistency in feature representations across diverse input conditions by measuring the distinction between original features and their reconstructions. Dense triplet loss ($\mathcal{L}_{dt}$) prioritizes significant pixel regions between modalities, thereby boosting model performance and generalization. $\mathcal{L}_{rtcl}$ is the recognizable target centroid loss, which is introduced in Section 3.2. Finally, λ_1 follows the baseline is set to 0.5 and λ_2 is set to 0.9.

4 Experimental Results and Analysis

4.1 Experimental Setting

Datasets

In this work, we adhere to established methodologies and perform experiments on the following publicly available datasets:

- Market-Sketch-1K [6] dataset, is derived from the Market-1501 dataset, which includes 498 identities from both the training and query sets. It contains 4,763 sketches from 996 identities and 32,668 photos from 1,501 identities, created by six artists with distinct styles, reflecting their subjective perceptions.

- PKU Sketch Re-ID [3] dataset features 200 individuals, each represented by one expertly hand-drawn sketch and two color photos, totaling 600 images. Five artists, each with different styles, created the sketches based on descriptions from voluntary witnesses, highlighting challenges in subjective perception and human posture despite the dataset's smaller size.

Dataset Segmentation

The PKU Sketch Re-ID dataset is divided into training and test sets. The training set consists of 150 sketches and 300 corresponding photos, with a total of 150 labels. The test set includes 50 sketches and 100 photos, totaling 50 labels, as detailed in Table1.

The sketches are divided into five distinct subsets (A, B, C, D, and E), each representing a different sketch style or source. In the train set, subsets A, B, C, D, and E contain 34, 15, 60, 25, and 16 sketches, separately. Similarly, the test set is made up of 12 sketches from subset A, 5 from B, 19 from C, 8 from D, and 6 from E, totaling 50 sketches. This strategic division ensures dataset balance, challenges the model's ability to generalize across various sketch styles, and enhances robustness in practical applications. The MARKET-SKETCH-1K dataset follows the already segmented rule.

Table 1. Details of the division of the PKU sketch Re-ID dataset.

Item	Sketch						Photos
	A	B	C	D	E	Total	
Train	34	15	60	25	16	150	300
Test	12	5	19	8	6	50	100

Evaluation metrics

For evaluation metrics, the performance is evaluated quantitatively by mean average precision (mAP) and cumulative matching characteristic (CMC) at Rank-1 (R-1), Rank-5 (R-5), and Rank-10 (R-10) .

Implementation details

The proposed model is implemented in PyTorch using an NVIDIA RTX3090 GPU, with ResNet-50 as the backbone pre-trained on ImageNet. All images are resized to 288×144, employing horizontal flipping and random erasing for data augmentation. For both datasets, mini-batches contain 8 randomly selected identities with 4 images each, totaling 32 images. The MARKET-SKETCH-1K dataset was trained for 80 epochs using the SGD optimizer, taking approximately 30 to 75 minutes. The learning rate starts at 1×10^{-2}, peaks at 9×10^{-2} during the first 9 epochs, and then adjusts as follows: 1×10^{-1} for epochs 10 to 19, 1×10^{-2} for epochs 20 to 49, and 1×10^{-3} for epochs 50 to 80. For the PKU Sketch Re-ID dataset, the SGD optimizer uses a learning rate of 9×10^{-3}, and training lasts about 9 minutes over 100 epochs. Different random seed values are set for stability: 0 for MARKET-SKETCH-1K and 42 for PKU Sketch Re-ID.

4.2 Results

Comparison results on the MARKET- SKETCH-1K dataset.

Table2 shows that our model sets a new state of the art under the S_1 single-query protocol, achieving 25.20% mAP and 24.70 % Rank-1. Notably, it outperforms CMAlign [16] and the latest baseline [6] by a clear margin, demonstrating its superior ability to handle sketches drawn by the same artist. Compared to AGW [17] with attribute features, our method gains substantially in both retrieval accuracy and matching precision, indicating that FFAM's feature disentanglement and RTCL's consistency enforcement yield more robust representations than relying on auxiliary attributes. Overall, these results confirm that CMDSL most effectively captures the consistent identity cues present in artist-specific sketches.

Table 2. Results on MARKET-SKETCH-1K dataset under the single-query S_1 setting. S_1 denotes all sketches made by the first artist. †attr denotes the utilization of attribute features for querying. Best with red color.

Method	Reference	Train	Query	mAP	R-1
AGW[16]	$CVPR'16$	†attr	attr	20.37	14.26
AGW[16]	$CVPR'16$	S_1	S_1	18.84	15.06
CMAlign[17]	$ICCV'21$			19.28	20.93
Baseline[6]	$MM'23$			22.89	22.63
Ours				25.20	24.70

Table 3. Comparison with other methods using allsketches from the MARKET-SKETCH-1K dataset. B denotes baseline, S denotes single-query and M denotes multi-query. Best with red color.

Method	Reference	mAP	R-1	R-5	R-10
DDAG [18]	$ECCV'20$	12.13	11.22	25.40	35.02
CMNAS [19]	$ICCV'21$	0.82	0.70	2.00	3.90
CAJ [20]	$ICCV'21$	2.38	1.48	3.97	7.34
MMN [21]	$MM'21$	10.41	9.32	21.98	29.58
DART [22]	$CVPR'22$	7.77	6.58	16.75	23.42
DCLNet [23]	$MM'22$	13.45	12.24	29.20	39.58
DSCNet [24]	$TIFS'22$	14.73	13.84	30.55	40.34
DEEN [25]	$CVPR'23$	12.62	12.11	25.44	30.94
B (S) [6]	$MM'23$	19.61	18.10	38.95	50.75
Ours (S)		20.87	17.34	41.60	54.35
DNS(S) [26]	$ECCV'24$	23.71	22.74	44.51	56.37
B (M) [6]	$MM'23$	24.45	24.70	50.40	63.45
Ours (M)		27.43	26.71	51.41	63.86

Table 3 presents a comprehensive comparison on MARKET-SKETCH-1K under both single-query (S) and multi-query (M) protocols. In the S setting, our CMDSL model achieves an mAP of 20.87 % and Rank-1 accuracy of 17.34 %, improving over the baseline's [6] 19.61 % mAP by 1.26 %, and outperforming other methods such as DCLNet [23] (13.45 % mAP) and DSCNet [24] (14.73 % mAP). However, DNS [26], an advanced approach published at ECCV24, still leads with 23.71% mAP and 22.74% R-1, indicating that single-query sketch retrieval remains challenging for our framework. Under the M protocol, which aggregates multiple sketch queries per identity, our model's performance increases dramatically: we obtain 27.43 % mAP and 26.71 % Rank-1, surpassing the baseline (M) 24.45 % mAP by 2.98 % and 24.70 % R-1 by 2.01 %. Moreover, Our multi-query mAP of 27.43% is 3.72% higher than DNS's 23.71%, and our Rank-1 of 26.71% beats their 22.74% by 3.97%. This strong improvement shows that CMDSL can integrate complementary subjective cues from multiple sketches. In summary, while DNS still excels in single-query, our approach demonstrates superior scalability and robustness in the multi-query setting, validating the design of CMDSL for scenarios where multiple witness sketches are available.

Comparison results on the PKU Sketch Re-ID dataset.

Table 4. Comparison with other methods using all sketches from the PKU Sketch Re-ID dataset. Best with red color.

Method	Reference	mAP	R-1	R-5	R-10	R-20
TripletSN [27]	$CVPR'16$	-	9.00	26.80	42.20	65.20
GN Siamese [28]	$TOG'16$	-	28.90	54.00	62.40	78.20
AFL-Net [3]	$MM'18$	-	34.00	56.30	72.50	84.70
RCD [29]	$CRR'21$	-	42.50	70.00	87.50	-
MDFL-Net [30]	$Neuro'20$	-	49.00	70.40	80.20	92.00
UFE [31]	$TMM'20$	-	57.14	79.59	89.80	93.88
CDA [32]	$TIFS'22$	-	60.80	80.60	88.80	95.00
UNIREID [33]	$CVPR'23$	-	69.80	88.60	95.80	-
Baseline [6]	$MM'23$	66.37	70.00	-	-	-
DCFF [34]	$JVCIR'24$	-	76.60	93.40	97.60	100.00
Ours		76.13	78.00	92.00	98.00	100.00

The comparison results are shown in Table 4, where we evaluated our method against several state-of-the-art approaches using the PKU Sketch Re-ID dataset. Triplet SN [27], designed for freehand sketches, achieved a Rank-1 accuracy of 9.00%, highlighting the challenges in extracting invariant features. GN Siamese [28], with dual GoogLeNet branches for Siamese and classification loss, improved this to 28.90%. AFL-Net [3], a cross-domain adversarial model, reached 34.00%, showing its ability to learn the identity and domain-invariant features. RCD [29], using random transformations, achieved a Rank-1 of 42.50%. MDFL-Net [30] attained a competitive Rank-1 accuracy of 49.00% by fusing multi-level features. The UFE [31] methods achieved a Rank-1 accuracy of 57.14% using unbiased feature extractors, while CDA [32] proposed an inter-domain attention mechanism, reaching 60.80%. UNIREID [33] attempts to leverage descriptive queries to study modality-agnostic person re-identification, achieving a rank-1 accuracy of 69.80%, while DCFF [34] proposes a cross-modal feature fusion network, attaining a rank-1 accuracy of 76.60%. The proposed method achieved 78.00% Rank-1 and 76.13% mAP. These results demonstrate that CMDSL generalizes robustly across datasets: despite the varied sketch styles and scene complexities of the PKU benchmarks, our FFAM and RTCL modules consistently extract dependable subjective features and enforce discriminative alignment, yielding stable and high-accuracy retrieval.

Evaluating on unignorable styles

Table 5 demonstrates that models trained on a single artist's sketches generalize poorly to unseen drawing styles. As the number of artists in the training set increases, performance on unseen-artist splits consistently improves, confirming that style diversity helps FFAM and RTCL learn robust, modality-invariant features. Moreover, when both training and evaluation employ multi-query sketches, mAP rises further, indicating that aggregating multiple queries yields more stable retrieval. Notably, even when training and testing protocols do not match (e.g., multi-query training with single-query testing), results still exceed those of single-artist baselines. These observations validate that greater artist diversity and query redundancy systematically strengthen the dependability of subjective features and enhance cross-artist generalization.

Table 5. Evaluation of unignorable styles on the MARKET-SKETCH-1K dataset during training and testing.

Test Train	S_6	$S_{5,6}$	$S_{4\ldots6}$	$S_{3\ldots6}$	$S_{2\ldots6}$
(a) mAP with single-query training and testing					
S_1	7.78	5.19	5.28	5.25	6.39
$S_{1,2}$	8.67	7.52	7.84	8.75	-
$S_{1\ldots3}$	13.85	11.58	10.92	-	-
$S_{1\ldots4}$	14.49	12.27	-	-	-
$S_{1\ldots5}$	16.11	-	-	-	-
(b) mAP with multi-query training and testing					
$S_{1,2}$	-	8.63	8.65	9.39	-
$S_{1\ldots3}$	-	10.82	11.49	-	-
$S_{1\ldots4}$	-	12.56	-	-	-
(c) mAP with single-query training and multi-query testing					
S_1	-	6.67	5.26	4.39	4.13
$S_{1,2}$	-	6.25	5.81	4.49	-
$S_{1\ldots3}$	-	9.19	6.78	-	-
$S_{1\ldots4}$	-	11.71	-	-	-
(d) mAP with multi-query training and single-query testing					
$S_{1,2}$	9.18	6.70	6.93	6.88	-
$S_{1\ldots3}$	9.00	7.56	7.43	-	-
$S_{1\ldots4}$	8.89	7.30	-	-	-
$S_{1\ldots5}$	9.11	-	-	-	-

4.3 Ablation studies

Efficacy of every element

We evaluate the effectiveness of each component on the MARKET-SKETCH-1K dataset. Each component is added independently to observe its individual contribution, as shown in Table 6. The results indicate that all components are beneficial, with FFAM achieving notable performance improvements. This further demonstrates the effectiveness of FFAM in enhancing the model's ability to extract dependable subjective features while suppressing modality-specific noise. Comparing the second row with the fourth row highlights the necessity of incorporating RTCL, and also verifies the effectiveness of our RTCL module.

Table 6. Ablation study of every element.Training and testing follow the multi-query protocol on MARKET-SKETCH-1K dataset.

Baseline	FFAM	RTCL	mAP	R-1
√	-	-	24.45	24.70
√	√	-	26.02	25.90
√	-	√	25.63	25.26
√	√	√	**27.43**	**26.71**

Attributes' noise studies.

Table 7. Ablation study of Attributes' noise. Training and testing follow the multi-query protocol on MARKET-SKETCH-1K dataset.

Noise coverage	**Baseline**		**Ours**	
	mAP	Rank-1	mAP	Rank-1
0/27	24.45	24.70	**27.43**	**26.71**
3/27	21.29	20.28	26.77	25.50
6/27	20.89	19.28	26.68	24.50
9/27	20.56	16.67	26.51	23.90

In the MARKET-SKETCH-1K dataset, each identity has 27 attribute labels. We conducted experiments by randomly introducing noise into some of these attributes (Table7). Despite increasing noise levels, this work's model maintained remarkable robustness, experiencing only a minor performance decline compared to other methods. This underscores the exceptional resilience and reliability of the proposed model when addressing attribute label noise.

The sensitivity of the hyperparameters λ_2 analysis

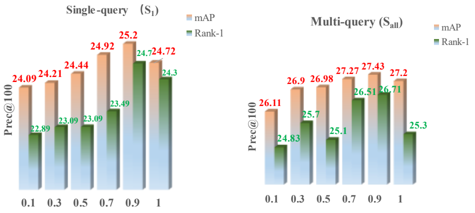

Fig. 3. Impact of λ_2 on MARKET-SKETCH-1K dataset.

The sensitivity of the hyperparameters λ_2 was analyzed to investigate its impact on the performance of the model. Fig. 3 present the results of the parameter tuning for λ_2

under both single-query and multi-query settings, respectively. These line plots illustrate how varying values of λ_2 affect the overall performance metrics, providing insights into the optimal selection of this hyperparameter for different retrieval scenarios. We can see that the performance peaks when the value of λ_2 is set to 0.9.

4.4 Visual Analysis

We employ Grad-CAM (Gradient-weighted Class Activation Mapping) [35] to highlight the regions of the images that are most influential in the decision-making process of our model after the FFAM. By doing so, we can see which parts of the sketch and photo pairs the model focuses on when determining a match.

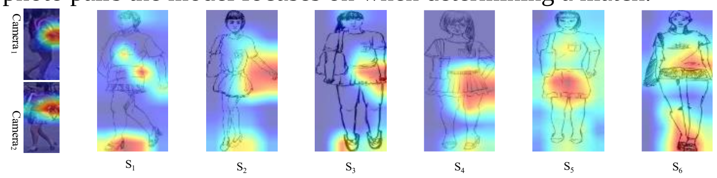

Fig. 4. Attention visualization of image pairs in MARKET-SKETCH-1K dataset.

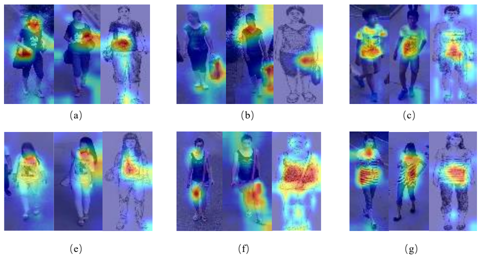

Fig. 5. Attention visualization of image pairs in PKU Sketch Re-ID dataset.

MARKET-SKETCH-1K dataset

We selected a pair of images in the MARKET-SKETCH-1K dataset to generate heatmaps containing two visable-images and six sketches, as shown in Fig. 4. In the heatmap, we can observe that our model consistently focuses on the same areas in the RGB images, indicating its ability to identify key features. However, due to the inherent differences between the sketch and RGB domains, the model's attention shifts when processing sketch images. While it maintains a focus on similar areas, such as the waist of the person in all images, it also extends its attention to additional parts of the sketches. This can be seen in images S1-S6, where the model not only highlights the waist but also pays considerable attention to the feet. This behavior suggests that the model is attempting to reconcile the stylistic differences between the sketches and the RGB images by expanding its focus to capture more diverse features, capturing dependable subjective features for subsequent RTCL. The retrieval results are shown in Fig. 6. We compare the our model with the baseline. As shown in Fig. 6(a), the correct person in the gallery can be found in the range of Rank-1 to Rank-3 with our model.

But the result in Fig. 6(b) shows that the baseline can only find the correct person in the range of Rank-1 to Rank-5.

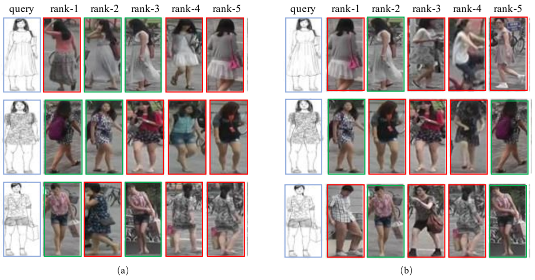

Fig. 6. Retrieval results on the Market-Sketch-1K dataset. (a) the complete model(b) the baseline model.

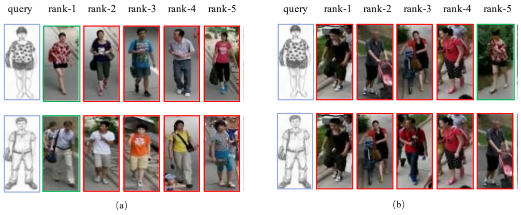

Fig. 7. Retrieval results on the PKU Sketch Re-id dataset. (a) the complete model(b) the baseline model.

PKU Sketch Re-ID dataset

We selected six pairs of images in the PKU Sketch Re-ID dataset to generate heatmaps containing twelve photographs and six sketches, as shown in Fig. 5. In this heatmap, we observe that in image pairs (a), (b), and (f), the attention areas are only partially similar. This partial overlap suggests that the model is starting to grasp the shared features between sketches and RGB images but still faces challenges due to domain differences. However, in image pairs (d), (e), and (g), the attention areas are largely similar, indicating a more robust alignment between the modalities. By reinforcing attention on dependable subjective cues, RTCL significantly enhances the model's ability to bridge the domain gap between sketches and RGB images. The retrieval results are shown in Fig. 7. We compare our model with the baseline. As shown in Fig. 7(a), our model can lock the correct pedestrian in the range of Rank-1 to Rank-3.But the result in Fig. 7(b) shows that the baseline only barely matches at Rank-5 or fails completely.

5 Conclusions and Limitations

Conclusions: This paper addresses the challenge of subjective variability in sketch-based person re-identification by proposing a novel Cross-Modal Dependable

Subjective Learning (CMDSL)framework. CMDSL is designed to extract and align reliable subjective cues across heterogeneous modalities, thereby enhancing shared representations. The core components of CMDSL are a dual-stream backbone, the Flexible Feature Aggregation Module (FFAM), and the Recognisable Target Centroid Loss (RTCL). Together, these modules work in concert to deliver significant performance gains. The robustness and generalization of this method are validated on two benchmark datasets, demonstrating its effectiveness in sketch re-identification tasks.

Limitations: The proposed CMDSL has achieved good performance. Due to the lack of attention to spatial image spatial information (Relative position of pedestrians between local areas), there is still room for further improvement. Of course, it is also the focus of our future work and research.

6 Acknowledgments

This research was funded by the Hubei Province Natural Science Foundation of China under Grant (No. 2023AFB357, 2021CFB568) and the Naval University of Engineering Foundation under Grant 2023508020.

References

1. Qingsong Hu, Huafeng Li, Zhanxuan Hu, and Feiping Nie. Diverse semantic information fusion for unsupervised person re-identification. Information Fusion, page 102319, 2024.
2. Shuang Li, Tong Xiao, Hongsheng Li, Bolei Zhou, Dayu Yue, and Xiaogang Wang. Person search with natural language description. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1970–1979, 2017.
3. Lu Pang, Yaowei Wang, Yi-Zhe Song, Tiejun Huang, and Yonghong Tian. Cross-domain adversarial feature learning for sketch re-identification. In Proceedings of the ACM International Conference on Multimedia, pages 609–617, 2018.
4. Rogerio Feris, Russel Bobbitt, Lisa Brown, and Sharath Pankanti. Attribute-based people search: Lessons learnt from a practical surveillance system. In Proceedings of IEEE International Conference on Multimedia Retrieval, pages 153–160, 2014.
5. Wenbin Zhang, Zhaoyang Li, Haishun Du, Jiangang Tong, and Zhihua Liu. Dual-stream feature fusion network for person re-identification. Engineering Applications of Artificial Intelligence, 131:107888, 2024.
6. Kejun Lin, Zhixiang Wang, Zheng Wang, Yinqiang Zheng, and Shin'ichi Satoh. Beyond domain gap: Exploiting subjectivity in sketch-based person retrieval. In Proceedings of the ACM International Conference on Multimedia, pages 2078–2089,2023.
7. Sergey Zagoruyko and Nikos Komodakis. Learning to compare image patches via convolutional neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 4353–4361, 2015.
8. Hao Ren, Ziqiang Zheng, and Hong Lu. Energy-guided feature fusion for zero-shot sketch-based image retrieval. Neural Processing Letters, 54(6):5711–5720, 2022.
9. Feifei Zhang, Sijia Qu, Fan Shi, and Changsheng Xu. Overcoming the pitfalls of vision-language model for image-text retrieval. In Proceedings of the ACM International Conference on Multimedia, pages 2350–2359, 2024.

10. Yi Bin, Haoxuan Li, Yahui Xu, Xing Xu, Yang Yang, and Heng Tao Shen. Unifying two-stream encoders with transformers for cross-modal retrieval. In Proceedings of the ACM International Conference on Multimedia, pages 3041–3050, 2023.
11. Lin Wu, Deyin Liu, Wenying Zhang, Dapeng Chen, Zongyuan Ge, Farid Boussaid, Mohammed Bennamoun, and Jialie Shen. Pseudo-pair based self-similarity learning for unsupervised person re-identification. IEEE Transactions on Image Processing, 31:4803–4816, 2022.
12. Cheng Deng, Xinxun Xu, Hao Wang, Muli Yang, and Dacheng Tao. Progressive cross-modal semantic network for zero-shot sketch-based image retrieval. IEEE Transactions on Image Processing, 29:8892–8902, 2020.
13. Deyin Liu, Lin Wu, Richang Hong, Zongyuan Ge, Jialie Shen, Farid Boussaid, and Mohammed Bennamoun. Generative metric learning for adversarially robust open-world person re-identification. ACM Transactions on Multimedia Computing, Communications and Applications, 19(1):1–19, 2023.
14. Jieru Jia, Qiuqi Ruan, and Timothy M Hospedales. Frustratingly easy person re-identification: Generalizing person re-id in practice. arXiv preprint arXiv:1905.03422, 2019.
15. Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 7132–7141, 2018.
16. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 770–778, 2016.
17. Hyunjong Park, Sanghoon Lee, Junghyup Lee, and Bumsub Ham. Learning by aligning: Visible-infrared person re-identification using cross-modal correspondences. In Proceedings of the IEEE International Conference on Computer Vision, pages 12046–12055, 2021.
18. Mang Ye, Jianbing Shen, David J. Crandall, Ling Shao, and Jiebo Luo. Dynamicdual-attentive aggregation learning for visible-infrared person re-identification. In Proceedings of the Springer European Conference on Computer Vision, pages 229–247. Springer, 2020.
19. Chaoyou Fu, Yibo Hu, Xiang Wu, Hailin Shi, Tao Mei, and Ran He. Cm-nas: Cross-modality neural architecture search for visible-infrared person re-identification.In Proceedings of the IEEE International Conference on Computer Vision, pages11823–11832, 2021.
20. Mang Ye, Weijian Ruan, Bo Du, and Mike Zheng Shou. Channel augmented joint learning for visible-infrared recognition. In Proceedings of the IEEE International Conference on Computer Vision, pages 13567–13576, 2021.
21. Yukang Zhang, Yan Yan, Yang Lu, and Hanzi Wang. Towards a unified middle modality learning for visible-infrared person re-identification. In Proceedings of the ACM International Conference on Multimedia, pages 788–796, 2021.
22. Mouxing Yang, Zhenyu Huang, Peng Hu, Taihao Li, Jiancheng Lv, and Xi Peng. Learning with twin noisy labels for visible-infrared person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,pages 14308–14317, 2022.
23. Sun, H., Liu, J., Zhang, Z., Wang, C., Qu, Y., Xie, Y., Ma, L.: Not all pixels are matched: Dense contrastive learning for cross-modality person re-identification. In: Proceedings of the ACM international conference on multimedia. pp. 5333–5341 (2022).
24. Zhang, Y., Kang, Y., Zhao, S., Shen, J.: Dual-semantic consistency learning for visible-infrared person re-identification. IEEE Transactions on Information Forensics and Security. 18, 1554–1565 (2022).
25. Yukang Zhang and Hanzi Wang. Diverse embedding expansion network and low-light cross-modality benchmark for visible-infrared person re-identification. In Pro-ceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2153–2162, 2023.

26. Yan Jiang, Xu Cheng, Hao Yu, Xingyu Liu, Haoyu Chen, and Guoying Zhao. Domain shifting: A generalized solution for heterogeneous cross-modality person re-identification. In European Conference on Computer Vision, pages 289–306. Springer, 2024.
27. Qian Yu, Feng Liu, Yi-Zhe Song, Tao Xiang, Timothy M Hospedales, and Chen-Change Loy. Sketch me that shoe. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 799–807, 2016.
28. Patsorn Sangkloy, Nathan Burnell, Cusuh Ham, and James Hays. The sketchy database: learning to retrieve badly drawn bunnies. ACM Transactions on Graphics, 35(4):1–12, 2016.
29. Yunpeng Gong, Liqing Huang, and Lifei Chen. Eliminate deviation with deviation for data augmentation and a general multi-modal data learning method. Computing Research Repository, abs/2101.08533, 2021.
30. Shaojun Gui, Yu Zhu, Xiangxiang Qin, and Xiaofeng Ling. Learning multi-level domain invariant features for sketch re-identification. Neurocomputing, 403:294–303, 2020.
31. Fan Yang, Yang Wu, Zheng Wang, Xiang Li, Sakriani Sakti, and Satoshi Nakamura. Instance-level heterogeneous domain adaptation for limited-labeled sketch-to-photo retrieval. IEEE Transactions on Multimedia, 23:2347–2360, 2020.
32. Fengyao Zhu, Yu Zhu, Xiaoben Jiang, and Jiongyao Ye. Cross-domain attention and center loss for sketch re-identification. IEEE Transactions on Information Forensics and Security, 17:3421–3432, 2022.
33. Cuiqun Chen, Mang Ye, and Ding Jiang. Towards modality-agnostic person re-identification with descriptive query. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 15128–15137, 2023.
34. Yu Ye, Jun Chen, Zhihong Sun, and Mithun Mukherjee. Data compensation and feature fusion for sketch based person retrieval. Journal of Visual Communication and Image Representation, 104:104287, 2024.
35. Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision, pages 618–626, 2017.