TriReID: Towards Multi-Modal Person Re-Identification via Descriptive Fusion Model

Yajing Zhai*
College of Computer Science and Electronic Engineering, Hunan University
Changsha, China
yajingzhai9@gmail.com

Yawen Zeng*
College of Computer Science and Electronic Engineering, Hunan University
Changsha, China
yawenzeng11@gmail.com

Da Cao†
College of Computer Science and Electronic Engineering, Hunan University
Changsha, China
caoda0721@gmail.com

Shaofei Lu†
College of Computer Science and Electronic Engineering, Hunan University
Changsha, China
sflu@hnu.edu.cn

ABSTRACT

The cross-modal person re-identification (ReID) aims to retrieve one person from one modality to the other single modality, such as text-based and sketch-based ReID tasks. However, for these different modalities of describing a person, combining multiple aspects can obviously make full use of complementary information and improve the identification performance. Therefore, to explore how to comprehensively consider multi-modal information, we advance a novel multi-modal person re-identification task, which utilizes both text and sketch as a descriptive query to retrieve desired images. In fact, the textual description and the visual description are understood together to retrieve the person in the database to be more aligned with real-world scenarios, which is promising but seldom considered. Besides, based on an existing sketch-based ReID dataset, we construct a new dataset, TriReID, to support this challenging task in a semi-automated way. Particularly, we implement an image captioning model under the active learning paradigm to generate sentences suitable for ReID, in which the quality scores of the three levels are customized. Moreover, we propose a novel framework named Descriptive Fusion Model (DFM) to solve the multi-modal ReID issue. Specifically, we first develop a flexible descriptive embedding function to fuse the text and sketch modalities. Further, the fused descriptive semantic feature is jointly optimized under the generative adversarial paradigm to mitigate the cross-modal semantic gap. Extensive experiments on the TriReID dataset demonstrate the effectiveness and rationality of our proposed solution.

*Both authors contributed equally to this research.
†Corresponding authors.

ICMR '22, June 27–30, 2022, Newark, NJ, USA

ACM ISBN 978-1-4503-9238-9/22/06…$15.00
https://doi.org/10.1145/3512527.3531397

CCS CONCEPTS

• **Information systems** → **Information retrieval**; • **Computing methodologies** → **Artificial intelligence**.

KEYWORDS

Multi-Modal Re-Identification, Descriptive Embedding, Generative Adversarial Networks, Multi-Modal Retrieval

ACM Reference Format:
Yajing Zhai, Yawen Zeng, Da Cao, and Shaofei Lu. 2022. TriReID: Towards Multi-Modal Person Re-Identification via Descriptive Fusion Model. In *Proceedings of the 2022 International Conference on Multimedia Retrieval (ICMR '22), June 27–30, 2022, Newark, NJ, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3512527.3531397

1 INTRODUCTION

With the flourishing of intelligent video surveillance equipment and its wide application in criminal investigation, how to quickly and accurately perform person re-identification (ReID) from massive cross-camera videos has been a key research question in both academia and industry. To overcome the limitation of the traditional image-to-image ReID method that is only suitable for retrieving the same person from the RGB images across different cameras [11, 42], the task of cross-modal person re-identification is proposed to solve more complex scenarios [8, 31]. Among them, verbal texts and drawing sketches are the most common ways to describe a person, which is very convenient for practical applications such as searching for criminal suspects [36, 37]. Unfortunately, the advantages of text and sketch are not fully exploited.

The text-based ReID task mainly solves the modality gap between language and vision. However, how to bridge the gap between the abstract text and the concrete visual image is still a tricky issue [2, 4]. Moreover, for the sketch-based ReID task, sketch and image are both visual semantic expressions, but the sketch is drawn with the help of text that only contains texture information and is limited by the lack of information such as color, posture and other information [21]. In fact, the two descriptive modalities of text and sketch are complementary [26]. The sketch is a visual representation like a RGB image, which has more concrete information than abstract text. Meanwhile, the text has the ability to bring more details such

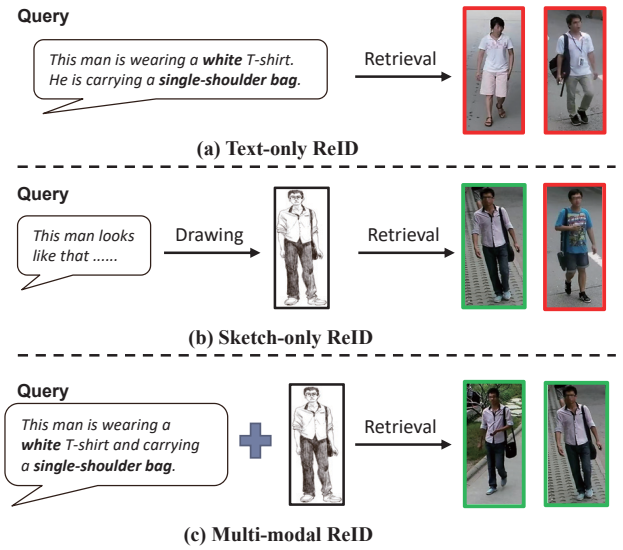


Figure 1: Examples of the retrieval query in the text-only ReID, sketch-only ReID and our multi-modal ReID. The textual description is expert in describing the appearance, while the sketch description conveys the details.

as gestures than the sketch. Therefore, how to combine the two descriptive information of sketch and text is worth investigating.

As revealed in Figure 1, there is a comparison for the cross-modal ReID methods based on the text-only query, sketch-only query, and multi-modal query, where red boxes represent negative images and green boxes indicate positive results. The text-only query in Figure 1 (a) contains detailed descriptions, such as "white", "single-shoulder bag", etc., but the retrieval result performs poorly due to the large semantic gap between textual and visual content. And as shown in Figure 1 (b), the sketch-only query is closer to the visual expression of RGB images to be retrieved, which has insufficient accuracy without detailed information. Therefore, we define a new multi-modal ReID task, which simultaneously utilizes text and sketch as a descriptive query to retrieve desired images. In Figure 1 (c), this multi-modal ReID solution shows excellent performance.

Although both text and sketch modalities are so important, the multi-modal person ReID task can not be carried out smoothly due to the lack of relevant datasets. Thus, to fill the vacuum of the dataset, we particularly construct a new dataset (**TriReID**) in the form of triplet, <text, sketch, RGB>, from an existing sketch-based dataset PKU [21]. Specifically, given a RGB image, we implement an image captioning model under the active learning paradigm to generate sentences suitable for ReID, in which the quality scores from three aspects are customized, i.e. fluency score, matching score, and diversity score. Moreover, we propose a novel framework named Descriptive Fusion Model (**DFM**) to investigate the multi-modal ReID comprehensively. In a detailed and exact way, we develop a flexible descriptive embedding module to fuse the text and sketch modalities as descriptive information. Thereafter, the descriptive feature and visual information are optimized for retrieval under the generative adversarial paradigm. Finally, the trained model is applied for person retrieval. By conducting experiments on the multi-modal person ReID dataset, we have explained the dominant position of our proposed DFM on overall performance comparison, ablation study, and micro-scope investigation.

The main contributions of this work are fourfold:

- To the best of our knowledge, this is the first work that considers two descriptive modalities as the query for multi-modal ReID, which fully explores the complementary semantics of text and sketch.
- We have collected a dataset to fill the vacuum of multi-modal ReID, which contains three modality information of textual description, drawing sketches, and RGB images about a same person.
- We contributed a DFM framework, which can integrate the descriptive modalities of text and sketch, and trained against each other with visual features to enhance the semantic understanding.
- Extensive experiments are performed to validate the effectiveness of our proposed DFM framework. Meanwhile, we release our source codes and dataset (The new dataset consists of the triplet <text, sketch, RGB>, the authors have only rights to release the textual description, please refer to the PKU Sketch ReID dataset[21] for the sketch and RGB images details.) to facilitate the research community[1].

2 RELATED WORK

In this paper, there are mainly two tasks that are tightly relevant: cross-modal person ReID and cross-modal retrieval. In this section, we will concisely introduce the related literature of these two parts.

2.1 Cross-Modal Person ReID

Existing methods of cross-modal person ReID can be classified into four categories: low resolution-based [20], infrared-based [5], text-based [36], and sketch-based [35] person ReID, which retrieve a person in one modality from the other modality. Li et al. [17] first studied the text-based cross-modal person ReID and proposed a method based on recurrent neural network and threshold neural attention mechanism. Li et al. [16] proposed to learn cross-modal embedding features and used a recurrent attention mechanism to refine the matching results. Liu et al. [19] introduced a graph attention convolutional network to learn the text and RGB image features, and adopted generative adversarial network to distinguish the two. Pang et al. [21] first proposed a cross-modal framework by adopting adversarial feature learning, in which the first ReID dataset containing sketch image is presented. Zhu et al.[45] adopted a Deep Surroundings-person Separation Learning (DSSL) model to effectively extract and match person information, and hence realize a higher retrieval accuracy.

Unfortunately, the advantages of text and sketch are not fully exploited. These works rarely consider the complementarity between text and sketch, resulting in sub-optimal performance. Therefore, how to take full advantage of both representation learning results from visual and descriptive information is worth investigating.

[1] https://icmr-2022.wixsite.com/trireid

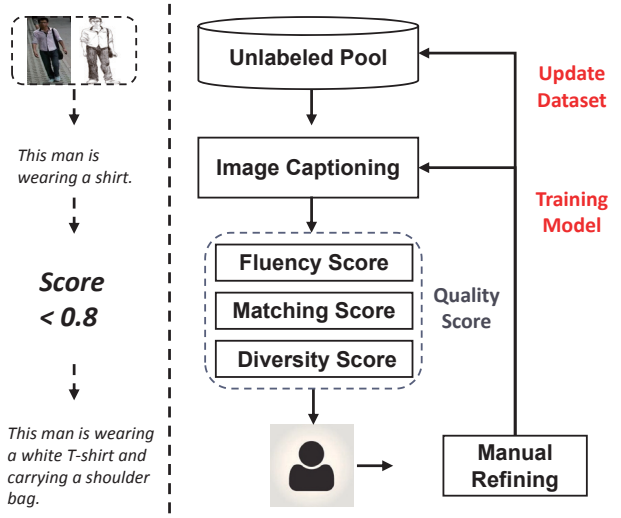


Figure 2: The process of automatic dataset construction for Multi-modal ReID task.

2.2 Cross-Modal Retrieval

The main effort on cross-modal retrieval is to eliminate the semantic gap between different modalities. Some studies [15, 40] have made great efforts to apply deep neural network for cross-modal retrieval. Recently, different from previous methods, GAN-based methods [12, 30, 46] have also been exploited to narrow the semantic gap [18]. Among pioneering works, Wang et al. [28] presented a cross-modal retrieval method based on adversarial learning to seek a useful common subspace. Peng and Qi [22] made use of GAN to successfully match up large-scale heterogeneous data. Dai et al. [6] handled the lack of insufficient discriminative information for different modalities via a generative adversarial network. Shao et al. [25] proposed a semi-supervised generative adversarial network for entity resolution. Zeng et al. [39] contributed a fresh solution to combine ranking and localization into a unified framework and the video moment retrieval task was thoroughly formulated as an adversarial learning problem. Qiang et al. [23] first calculated semantic similarity for cross-modal retrieval. And then an adversarial modality discriminator is constructed as a modality discriminator to calculate the similarity of each modality feature. Xu et al. [34] put forward a method for the scalability problem of cross-modal retrieval tasks, which simultaneously learned cross-modal network reconstruction and modal adversarial semantic association relations. Wu et al. [32] proposed a network guided by the adversarial scheme for cross-modal retrieval aiming to learn modality-specific features and the modality-shared features for each modality.

Inspired by the above work, we optimize the descriptive and visual features under the generative adversarial paradigm. Through this way, descriptive and visual features are learned from each other and the gap between different modalities is well reduced.

3 TRIREID DATATSET

3.1 Automatic Datatset Construction

Since there is no existing dataset containing the three modalities (i.e RGB, text and sketch), we contribute a new dataset, TriReID,

Table 1: Comparisons between different datasets. "-" indicates inapplicable.

Dataset	Text	Sketch	RGB	Sample
VIPeR [9]	-	-	1,264	1,264
Market-1501 [43]	-	-	32,668	32,668
PKU [21]	-	200	400	200
TriReID(Ours)	5,600	200	5,600	39,200

to facilitate the research community. Compared to collecting delicate and expensive sketches, we choose to automaticlly construct TriReID via adding textual description on the sketch-based dataset PKU [21]. Figure 2 illustrates the pipeline of our automatic dataset construction approach, which implements the image captioning under the active learning paradigm to generate sentences suitable for person ReID. Active learning is an interrogation method between the model and human experts, which reduces annotation workload. Furthermore, this automatic pipeline is divided into three steps, namely image captioning, automatic scoring, and expert refining.

1) Image Captioning. We adopt the image captioning model proposed in [27] for simple sentence generation. However, the sentences generated by the general models do not focus on describing person, so we design a quality score to constrain the generation process.

2) Automatic Scoring. The quality score $\mathcal{S}_{qua}$ is evaluated from three levels, i.e. fluency score $\mathcal{S}_{flu}$, matching score $\mathcal{S}_{mat}$, and diversity score $\mathcal{S}_{div}$, which is the score sum of adjective, noun and verb three kinds of words, i.e. adj., n. and v..

$$\mathcal{S}_{qua} = \lambda_1 \mathcal{S}_{flu} + \lambda_2 \mathcal{S}_{mat} + \lambda_3 \mathcal{S}_{div}, \tag{1}$$

where λ_1, λ_2, λ_3 denote the balance factors of three levels. Among them, the fluency score measures whether the sentence is fluent and reasonable, which is represented by the perplexity score (PPL). The matching score evaluates whether the sentence completely expresses the image, which matches the object similarity between the generated text and the image. And the diversity score calculates whether the sentence describes the attributes of pedestrians in ReID task, which highlights the adjectives or verbs in the sentence.

$$\mathcal{S}_{flu} = -\log[PPL(sentence)], \tag{2}$$

$$\mathcal{S}_{mat} = \underset{object_t}{MEAN}(\underset{object_v}{MAX}(Sim(object_t, object_v))), \tag{3}$$

$$\mathcal{S}_{div} = \sum sco_{word}, (word = n., adj., v.),$$
$$sco_{word} = \begin{cases} 1, cnt(word) > 3 \\ 0.8, cnt(word) = 3 \\ 0.5, cnt(word) = 2 \\ 0, cnt(word) = 1 \end{cases}. \tag{4}$$

where Sim represents inner-product similarity. sco_{word} is the score of adj., n. and v. respectively, cnt represents vocabulary count. Take the adjective as an example, if the number of adjectives in the sentence is greater than 3, we set the diversity score to 1. For the ReID task, we observe that more than 3 descriptive terms are sufficient for retrieval.

3) Expert Refining. For the triplets whose quality score is less than the threshold, they will be modified by human experts. The

Figure 3: The examples of TriReID dataset.

Table 2: The Statistical Analysis of the TriReID.

Item	Statistics
vocabulary size	1,400
word type	adj. v. n.
average length of sentences	22.52
minimum length of sentences	6
maximum length of sentences	58
frequent word	black

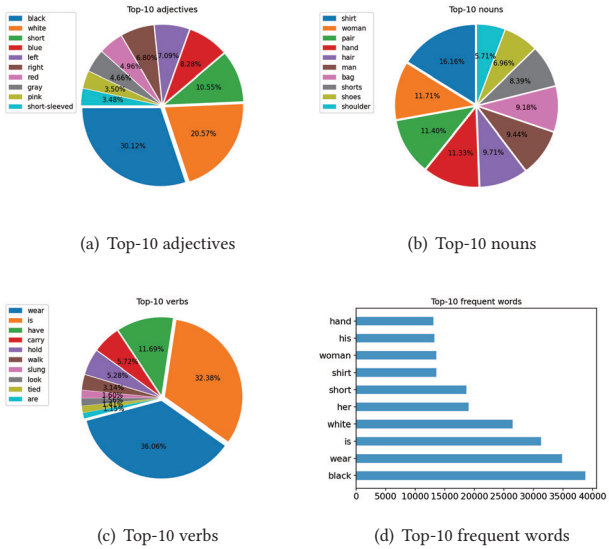
refined samples return to the dataset and participate in training until the quality of all annotations meets the requirements. Through this interactive active learning paradigm, the descriptive sentences of the ReID dataset are semi-automatically constructed. As much as possible, experts are asked to make sentences include more adjectives and nouns than verbs.

Moreover, to expand the TriReID dataset to a sufficient scale, we invite multiple experts to obtain various textual descriptions through the above process. Therefore, the image captioning model supervised by different experts produces diverse sentences. Furthermore, for RGB images, we also extend the simulation of better light, low resolution, and dark night scenes that are common in the real world. Main methods contain blur, brighter, horizontal, adding salt and grey, etc., data enhancement. Finally, we collect 39, 200 samples containing three modalities, as shown in Table 1. The TriReID dataset has the most comprehensive modalities and large-scale, suitable for research and application in both academia and industry. Furthermore, Figure 3 presents some examples of the TriReID dataset.

3.2 Dataset Analysis

To ensure the quality of the contributed dataset, we comprehensively analyze the TriReID dataset from property quality, diversity quality, and visualization quality.

Property Quality. The textual description statistical analysis of the TriReID dataset is shown in Table 2. According to statistics, there are mainly three types of words in the descriptional sentences, including adj., v., n., and the vocabulary size is 1,400. Besides, the



(a) Top-10 adjectives (b) Top-10 nouns

(c) Top-10 verbs (d) Top-10 frequent words

Figure 4: Diversity analysis

average length of sentences is 22.52 and the minimum length of the sentences is 6. And the most frequently occurring word in the sentence is "black".

Diversity Quality. To measure the sentence diversity of TriReID dataset, we count word frequency in all sentences. Figure 4 illustrates the frequency of top 10 in the textual description. We observe that the words with high frequency are mostly color, descriptive words, such as "blue", "bag", which shows that in ReID, detailed description is effective for retrieving.

Visualization Quality. To verify the rationality of the human expert's operation, we also check some cases as shown in the Figure 3. Specifically, we randomly select the captioning performance supervised by 7 experts. The annotations of different experts have different emphasis and the quality score plays a balance, which makes the sentences focus on the person itself as much as possible.

4 OUR PROPOSED FRAMEWORK

In this part, we elaborate on the detailed contents of our solution. The overall framework of our proposed DFM is depicted with an illustration in Figure 5. Generally speaking, it is composed of three components: descriptive semantic fusion, generative adversarial align, and person ReID. 1) Descriptive semantic fusion fuses data from multiple modalities to ensure the descriptive semantic space and visual semantic space reasonably; 2) Generative adversarial alignment aligns reconstructed paired-images by decomposing visual information and descriptive information into content information features and style information features, and then exchange them for further encoder images, respectively; and 3) Person ReID generates a descriptive-visual space feature representation for person retrieval.

4.1 Problem Formulation

Given the descriptive information containing both the textual description and sketch description, the task of multi-modal person

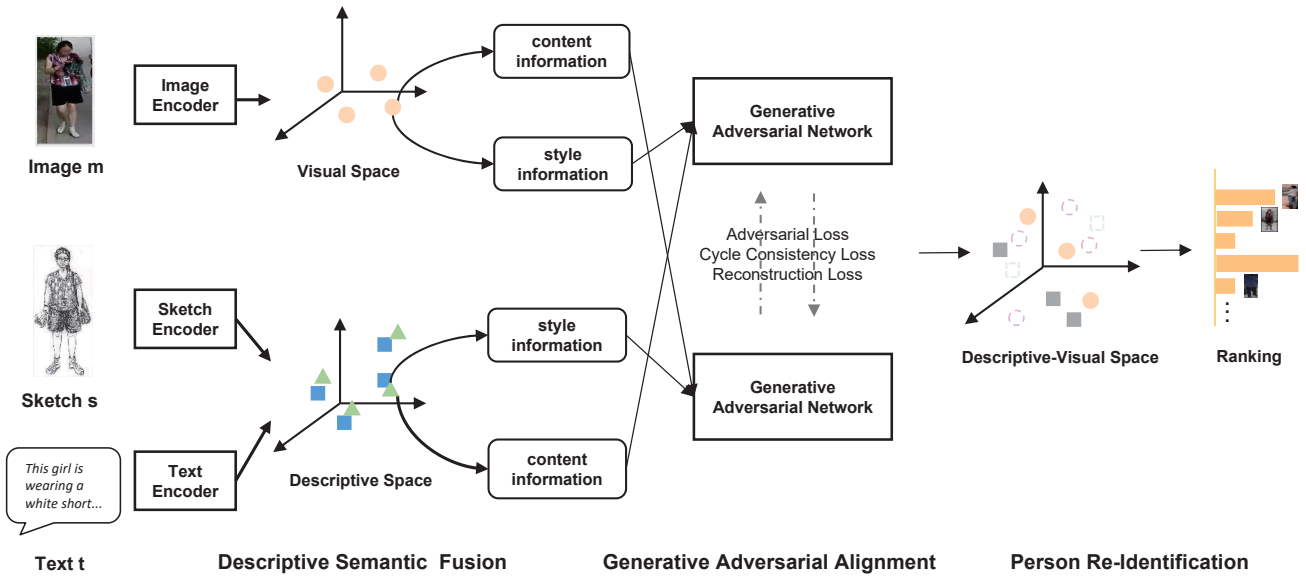


Figure 5: The graphical representation of our proposed DFM framework. The input is the three modalities of RGB image, text and sketch three modalities, and the output is the similarity between RGB and descriptive modalities (i.e. text and sketch).

ReID is to retrieve a list of relevant real RGB images. A sketch image denoted is as $S = \{s_n^N \mid s \in R^{H\times W\times 1}\}$, the corresponding textual descriptions are denoted by $T = \{t_n^N \mid t \in R\}$, and $M = \{m_n^N \mid m \in R^{H\times W\times 3}\}$ represents for the real RGB images, where N stands for the number of person identity, $n = \{1, 2, ..., N \mid n \in N\}$, H and W respectively represent the height and width of the RGB image. Each set of $<t_n, s_n, m_n>$corresponds to the same individual identity.

The features of sketch image and real image are extracted by a Resnet-50 [10] and denoted as f_v. It is worth noting that sketch and RGB share the same extractor since they are both visual representations. At the same time, the textual embedding is performed via the Doc2Vec [14] and is represented by f_t. Mathematically, the above process is expressed in formulas:

$$\begin{cases} Z_{\boldsymbol{m}} = f_v(m) \\ Z_{\boldsymbol{s}} = f_v(s) \\ Z_{\boldsymbol{t}} = f_t(t) \end{cases}, \tag{5}$$

where $Z_{\boldsymbol{m}}$, $Z_{\boldsymbol{s}}$, $Z_{\boldsymbol{t}}$ are extracted RGB features, sketch features and text features, respectively.

4.2 Descriptive Semantic Fusion

The first component of our DFM framework is descriptive semantic fusion, which highlights the descriptive semantic information. The sketch has both descriptive semantics similar to text and visual semantics consistent with the semantic expression of real RGB images. Since with the help of the sketch modality which is specially employed, this paper models a descriptive semantic space and a visual semantic space simultaneously.

Firstly, we extract the sketch, text and RGB image features, separately. And in the descriptive space, an attention pooling network is utilized to construct the relations between visual features and semantic embeddings, which fuses the representations of text and sketch. It passes through a Multi-Layer Perceptron at first and then is aggregated to represent the whole descriptive information Z_d, and adjusts the weights according to their importance. The descriptive semantic fusion is formally formulated as:

$$\begin{cases} \alpha = (Z_{\boldsymbol{t}} \cdot \boldsymbol{W}_1 + Z_{\boldsymbol{s}} \cdot \boldsymbol{W}_2)\, \boldsymbol{W}_3 \\ Z_{\boldsymbol{d}} = \mathrm{softmax}(\alpha) \cdot Z_{\boldsymbol{s}} \cdot \boldsymbol{W}_4 \end{cases}, \tag{6}$$

where $\boldsymbol{W}_1, \boldsymbol{W}_2, \boldsymbol{W}_3, \boldsymbol{W}_4$ are weight matrices. We compute a score α with weight vectors $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$ and normalize the scores with a softmax function, which is an experienced practice in previous work [1, 33]. Through this way, the descriptive information $\boldsymbol{Z_d}$ is fused via textual description and visual sketch.

4.3 Generative Adversarial Alignment

As shown in Figure 5, the second component of our DFM model is generative adversarial alignment, which eases the modality gap between descriptive feature $\boldsymbol{Z_d}$ and visual feature $\boldsymbol{Z_m}$. Inspired by the effective performance of adversarial learning [37, 44], we try to separate the style information and content information to find the commonality and individuality of different modalities under the adversarial paradigm. Specifically, the separation process is implemented by two generative adversarial networks under various losses.

This component includes two generators G_1,G_2 and two discriminators D_1,D_2. And the content features $\boldsymbol{Z_m^c}$,$\boldsymbol{Z_d^c}$ and style features $\boldsymbol{Z_m^s}$,$\boldsymbol{Z_d^s}$ are initialized from the original features $\boldsymbol{Z_m}$, $\boldsymbol{Z_d}$.

4.3.1 ***Adversarial Loss***. The adversarial loss $\mathcal{L}_{adv}$ swaps the content and style of the visual and descriptive modalities so that

Algorithm 1: DFM Framework

Input:
  Sketch: $S = \{s_n^N \mid s \in R^{H\times W\times 1}\}$
  Textual description: $T = \{t_n^N \mid t \in R\}$
  RGB image: $M = \{m_n^N \mid m \in R^{H\times W\times 3}\}$
Output: Rank-K retrieval results
DFM:
  // Feature extraction
  Extract features Z_m, Z_s, Z_t according to Eq.(5);
  // Descriptive information generation
  Generate descriptive feature Z_d using Eq.(6);
repeat
  for $(S, T, M) \in R$ **do**
    // Generator;
    Generate fake feature;
    separate descriptive feature to style and content;
    separate RGB image feature to style and content;
    // Discriminator;
    Top-K ranking retrieval results;
    Get the loss $\mathcal{L}$ in Eq.(7)-(10);
    Optimization.
  end
until *convergence*;

the individuality of modalities is highlighted:

$$\mathcal{L}_{adv} = \mathbb{E}[\log D_1(\boldsymbol{Z_m})] + \mathbb{E}[\log(1 - D_1(G_1(Z_{\boldsymbol{m}}^{\boldsymbol{c}}, Z_{\boldsymbol{d}}^{\boldsymbol{s}}))] + \mathbb{E}[\log D_2(\boldsymbol{Z_d})] + \mathbb{E}[\log(1 - D_2(G_2(Z_{\boldsymbol{d}}^{\boldsymbol{c}}, Z_{\boldsymbol{m}}^{\boldsymbol{s}}))]. \quad (7)$$

Among them, the discriminators maximize the formula, while the generators do the opposite operation to ensure that the translated image is similar to the real image in the visible domain.

4.3.2 ***Cycle Consistency Loss.*** Inspired by CycleGAN [29, 47], we employ a cycle-consistency loss to further ensure that the generated samples retain characteristics of the original samples. This loss $\mathcal{L}_{\text{cyc}}$ can be expressed as below,

$$\mathcal{L}_{cyc} = \mathbb{E}[||G_1(G_2(\boldsymbol{Z_m})) - \boldsymbol{Z_m}||] + \mathbb{E}[||G_2(G_1(\boldsymbol{Z_d})) - \boldsymbol{Z_d}||]. \quad (8)$$

4.3.3 ***Reconstruction Loss.*** Meanwhile, as for the reconstructed images, we formulate the reconstruction loss $\mathcal{L}_{\text{rec}}$ as below,

$$\mathcal{L}_{rec} = \mathbb{E}[||G_1(Z_{\boldsymbol{m}}^{\boldsymbol{c}}, Z_{\boldsymbol{m}}^{\boldsymbol{s}}) - \boldsymbol{Z_m}||] + \mathbb{E}[||G_2(Z_{\boldsymbol{d}}^{\boldsymbol{c}}, Z_{\boldsymbol{d}}^{\boldsymbol{s}}) - \boldsymbol{Z_d}||], \quad (9)$$

Under these three losses, the style information and content information are separated, which is beneficial for representation learning and improving the retrieval performance.

4.4 Person ReID

Through the above steps, we unify the descriptive and visual features into the descriptive visual space. Similarity learning is utilized to judge whether the features vectors belong to different persons or not. In the testing phase, we employ the Euclidean distance to calculate the distance score between descriptive content $Z_{\boldsymbol{d}}^{\boldsymbol{c}}$ and visual content $Z_{\boldsymbol{m}}^{\boldsymbol{c}}$. Therefore, a higher score will be generated for a positive pair of person images than those of negative counterparts.

Table 3: Comparison with the State-Of-The-Art Methods and Various Components Combinations in TriReID dataset.

Method	R@1	R@5	R@10
DCMP	0.039	0.117	0.132
DLCM	0.020	0.060	0.110
JSIA	0.020	0.100	0.180
DHLSVM	0.051	0.168	0.283
Triplet SN	0.090	0.268	0.422
DFM w/o GAN	0.020	0.014	0.300
concatenate	0.080	0.222	0.380
dot	0.080	0.220	0.301
DFM	**0.200**	**0.500**	**0.640**

Here, λ_{reid} means the person ReID loss, which loss function contains a cross-entropy loss $\mathcal{L}_{\text{cls}}$ and a triplet loss $\mathcal{L}_{\text{tri}}$. Furthermore, $\mathcal{L}_{\text{cls}}$ and $\mathcal{L}_{\text{tri}}$ are respectively employed for identity learning and reduction about the distances between images of the same person, and the distances between the different person.

Overall, the objective function of our method DFM can be denoted as:

$$\mathcal{L} = \lambda_{cyc}\mathcal{L}_{cyc} + \lambda_{adv}\mathcal{L}_{adv} + \lambda_{rec}\mathcal{L}_{rec} + \lambda_{reid}\left(\mathcal{L}_{cls} + \mathcal{L}_{\text{tri}}\right), \quad (10)$$

where λ_{cyc}, λ_{adv}, λ_{rec} and λ_{reid} are the balance factors of cycle-consistency loss, adversarial loss, reconstruction loss and person ReID loss respectively. The workflow of our proposed DFM is elaboratively illustrated in Algorithm 1.

5 EXPERIMENTS

There are extensive experiments conducted on TriReID dataset that have answered the following four research questions:

RQ1 How does the DFM proposal perform as compared to other state-of-the-art methods?

RQ2 How do components of DFM and fusion method promote the performance of DFM?

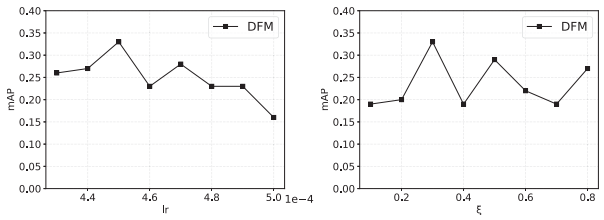
RQ3 How do different parameter settings (e.g., the margin, and the learning rate) have an effect on our framework?

RQ4 Can we visualize the generative images of our method?

5.1 Experimental Settings

5.1.1 ***Dataset.*** In this paper, we first divided the TriReID dataset into 3 non-overlapping sets, with 50%, 25% and 25% randomly selected person identities and their corresponding modalities (i.e. RGB, text and sketch) for training, validation and testing, respectively. The validation set and the training set make up the final training set and we performed the evaluation on the testing set.

5.1.2 ***Evaluation Protocols.*** We adopted the Rank-k and the mean Average Precision (mAP) to evaluate the performance [31]. Specifically, given a query textual description and a sketch, each method outputs prediction scores for all testing images in the gallery and ranks them accordingly. R@1, R@5, R@10 are reported for all our experiments on the dataset. The higher the value, the better the performance.



(a) Learning rate. (b) Margin.

Figure 6: Parameter tuning and convergence analysis in terms of mAP. The x-axis is the tuning range of learning rate and margin, while the y-axis is the mAP for performance evaluation.

5.1.3 ***Baselines.*** The effectiveness of our method is verified by comparing it with several state-of-the-art methods. Among them, these baselines can be classified into two cross-modal retrieval branches, Text-based ReID solutions, and Sketch-based ReID approaches.

- **DCMP** [41] uses feature projection operation, which includes CMPM loss and CMPC loss. The CMPM loss minimizes the Kullback-Leibler(KL) divergence of distributions. Furthermore, this method can categorize vector representations of the two modalities via the CMPC loss.
- **DLCM** [44] adopts the two-stream network structure as the backbone for cross-modality visible-infrared ReID. The first block is to capture modality-specific information for two modalities, while the remaining blocks are shared to learn modality sharable features.
- **JSIA** [29] performs set-level and instance-level introduced to gain cross-modal image features and generate some pairs of images. Moreover, the model performs directly alignment by calculating minimum distances of paired images.
- **DHLSVM** [21] contains HOG and LBP, which are classical features and assistant features for sketch recognition respectively. Then it is ranking scores by uniting rankSVM with the two features [21].
- **Triplet SN** [38] is a model mainly utilized for the research of sketch images. Following [38], after pre-training the model on edge-maps of ImageNet [7], this paper extracts features from sketch and RGB images and tests on Triplet SN [21].

5.1.4 ***Implementation and Hyper-Parameter Setting.*** We applied our method based on PyTorch framework[2] on a server equipped with a NVIDIA 2080TI-11G GPU. The feature generators were modified from ResNet [10]. Specifically, the employment of the mini-batch training is being widely used and the training dataset is split into small batches to reduce model costs and update model parameters.

The Adam optimizer [13] and back-propagation are mainly utilized to train our model and update the corresponding networks' parameters. Additionally, we performed data augmentation by cropping a 256×128 image as input and we performed the image flipping with 0.5 rate. For feature extraction, sketch image features and real image features are denoted as 2,048 dimensional vectors. And the dimensionality of the sentence feature embedding is 256.

Besides, we set the learning rate as 0.00045 and the batch size as 12. The cycle-consistency loss balance factor λ_{cyc} is set to 10, the adversarial loss balance factor λ_{adv} is set to 1, the person re-identification loss balance factor λ_{reid} and the reconstruction loss balance factor λ_{rec} are respectively set to 1. As for the triplet loss, we set the margin parameter ξ = 0.3. Specifically, we will comprehensively discuss parameter tuning in Section 5.4.

[2] http://www.pytorch.org

5.2 Overall Performance Comparison (RQ1)

To demonstrate the effectiveness of our proposed method, we compared it with several state-of-the-art approaches: 1) DCMP; 2) DLCM; 3) JSIA; 4) DHLSVM; 5) Triplet SN. All of these methods belong to cross-modal person ReID and we use our collected TriReID dataset to conduct experiments. Specifically, DCMP, DLCM, JSIA are classified as the cross-modal person ReID algorithms based on textual description query. And DHLSVM and Triplet SN belong to cross-modal person ReID methods based on sketch query.

Table 3 reveals the experimental results. There are several observations as follows: 1) Our DFM approach obtains a better performance than state-of-the-art baselines, which employs the adversarial learning paradigm between the visual information and the descriptive information to optimize the identification performance. It is obvious that the combination of text and sketch greatly improves the retrieval performance of multi-modal person ReID, which reaches 0.200 for R@1. 2) Cross-modal person ReID methods based on sketch query, DHLSVM and Triplet SN outperform cross-modal ReID retrieval approaches based on textual description query. This is probably because the sketch is highly related to the other two modalities, which is neglected in traditional cross-modal person ReID schemes.

5.3 Ablation Study (RQ2)

The overall performance comparison shows that DFM achieves the best performance, which proves the effectiveness of our proposed solution. To further distinguish the importance of GAN, distinct ways of fusion, we conducted the related ablation studies. First of all, DFM is compared with its variants, DFM w/o GAN, which represent DFM without GAN. In addition, we also compared our fusion method which adopted Multi-layer Perceptron (MLP) attention method with the other two fusion methods of concatenate and dot.

The performance of different component combinations is illustrated in Table 3, which illustrates the experimental results of the method of DFM and its simplified variants. The conclusions are threefold: 1) DFM meaningfully exceeds DFM w/o GAN. This displays that the GAN component is conducive to the multi-modal person ReID. The effectiveness of our DFM approach is certified, which indicates its efficiency in aggregating multiple components. 2) Jointly observing the performance of DFM, concatenate, dot, we can infer that textual description and sketch employ MLP attention fusion exceeds the concatenate and dot fusion method. According to the results, the attention mechanism fusion method achieves at least 12% improvement than the other two. This indicates that

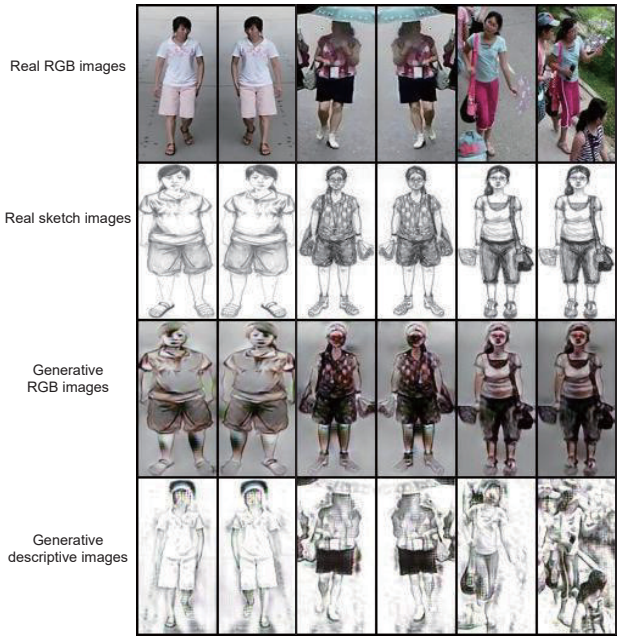


Figure 7: The visualization of generated images from our proposed framework

the component of MLP attention network we used is beneficial to the cross-modal person ReID. 3) The model has poor performance without using the GAN method. It reveals that the necessity of the components proposed in our framework.

5.4 Sensitivity of Parameters and Convergence Analysis (RQ3)

As we all know, the sensitivity of some factors has a crucial impact on the robustness and rationality of the model. Therefore, we investigate the impacts of several factors to prove the effectiveness of our proposed DFM framework, namely, the learning rate and the training margin ξ.

5.4.1 Impact of Learning Rate. Adam usually produces faster convergence and it adopts adaptive learning rates for different parameters to update the learning rate for each parameter. The parameter tuning results of learning rate are demonstrated in Figure 6(a). We use a step size of 0.00001 for fine-tuning and the results obtained have obvious changes. Among them, when the learning rate is 0.00045, the effect is the best.

5.4.2 Impact of Margin. The strategy of utilizing margin measures the gap between the positive pair and the negative pair. We reveal the performance of DFM w.r.t. different margin settings in Figure 6(b), which indicates the influence of margin for DFM. The values of mAP perform immediate changes along with the increasing of ξ, reaching its highest value when $\xi = 0.3$. The impact of margin has been proven by traditional retrieval algorithms [3, 24].

5.5 Visualization (RQ4)

The intention of this work is to learn the descriptive features combined with text and sketch, and visual features from the generative descriptive embedding network. To better understand our proposed DFM network, we exploited some visualization studies. More specifically, we randomly select descriptive features and visual features into the DFM framework. In addition, the RGB image and descriptive information of each person could separate style information and content information. Our generative module generates a new person descriptive image and RGB image by swapping the style features or content features of the two modal information respectively and minimizes their distances to reduce the cross-modality gap.

As Figure 7 shows, we display the cross-modality paired-images generated by our method. Among them, the first two rows are real RGB images and sketch images by professional artists, respectively. Then in the third row, we could see a fake RGB image generated by descriptive style features and RGB content features. Finally, the fourth row is generated fake descriptive images by encoding descriptive content features and RGB style features. As a result, our proposed DFM could generate features similar to RGB images to eliminate modal differences as much as possible by adversarial learning.

6 CONCLUSION AND FUTURE WORK

In this paper, we have addressed the new problem of the multimodal ReID, which utilizes text and sketch together as a descriptive query to retrieve desired images. To support this task, we construct the TriReID dataset based on an existing sketch dataset in a semi-automatic manner, which will be released publicly. We further propose a DFM framework, which fuses textual description and sketch information as descriptive feature, and then optimizes the descriptive space. Further, the descriptive features and visual information are jointly optimized for retrieval under the adversarial learning strategy. Experimental results demonstrate that our method achieves the best performance on a great margin.

In the future, we will expend this work from the following two aspects. First of all, the sketch only has the outline of the target person which is relatively insufficient as compared with the image, but the sketch features are highly related to the RGB image features. Therefore, some sketch information can be supplemented by generative model. In addition, we consider to apply graph representation learning to link the gap among different modalities via generating fine-grained semantic structured representations from images and texts of a person in the scene graphs.

ACKNOWLEDGEMENT

The authors are highly grateful to the anonymous referees for their careful reading and insightful comments. The work is supported by the National Natural Science Foundation of China (No. 61802121), the Natural Science Foundation of Hunan Province (No. 2019JJ50057), the Special Funds for the Construction of Innovative Provinces in Hunan Province of China (No. 2020SK2066), and the Fundamental Research Funds for the Central Universities.

REFERENCES

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural Machine Translation by Jointly Learning to Align and Translate. *arXiv preprint arXiv:1409.0473* (2014).

[2] Dapeng Chen, Hongsheng Li, Xihui Liu, Yantao Shen, Jing Shao, Zejian Yuan, and Xiaogang Wang. 2018. Improving Deep Visual Representation for Person Re-Identification by Global and Local Image-Language Association. In *Proceedings of the European Conference on Computer Vision*. 54–70.

[3] Jing-Jing Chen, Chong-Wah Ngo, Fu-Li Feng, and Tat-Seng Chua. 2018. Deep Understanding of Cooking Procedure for Cross-Modal Recipe Retrieval. In *Proceedings of the 26th ACM International Conference on Multimedia*. 1020–1028.

[4] Shiyu Chen, Yawen Zeng, Da Cao, and Shaofei Lu. 2022. Video-Guided Machine Translation via Dual-Level Back-Translation. *Knowledge-Based Systems* 245 (2022), 108598.

[5] Yehansen Chen, Lin Wan, Zhihang Li, Qianyan Jing, and Zongyuan Sun. 2021. Neural feature search for rgb-infrared person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 587–597.

[6] Pingyang Dai, Rongrong Ji, Haibin Wang, Qiong Wu, and Yuyu Huang. 2018. Cross-Modality Person Re-Identification with Generative Adversarial Training. In *International Joint Conference on Artificial Intelligence*, Vol. 1. 2.

[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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 248–255.

[8] Liying Gao, Kai Niu, Zehong Ma, Bingliang Jiao, Tonghao Tan, and Peng Wang. 2021. Text-Guided Visual Feature Refinement for Text-Based Person Search. In *Proceedings of the 2021 International Conference on Multimedia Retrieval*. 118–126.

[9] Douglas Gray, Shane Brennan, and Hai Tao. 2007. Evaluating Appearance Models for Recognition, Reacquisition, and Tracking. In *Proc. IEEE International Workshop on Performance Evaluation for Tracking and Surveillance*, Vol. 3. Citeseer, 1–7.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 770–778.

[11] Tianyu He, Xu Shen, Jianqiang Huang, Zhibo Chen, and Xian-Sheng Hua. 2021. Partial person re-identification with part-part correspondence learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9105–9115.

[12] Yiqi Jiang, Weihua Chen, Xiuyu Sun, Xiaoyu Shi, Fan Wang, and Hao Li. 2021. Exploring the Quality of GAN Generated Images for Person Re-Identification. In *Proceedings of the 29th ACM International Conference on Multimedia*. 4146–4155.

[13] Diederik P Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980* (2014).

[14] Quoc Le and Tomas Mikolov. 2014. Distributed Representations of Sentences and Documents. In *International Conference on Machine Learning*. PMLR, 1188–1196.

[15] Jinxing Li, Mu Li, Guangming Lu, Bob Zhang, Hongpeng Yin, and David Zhang. 2020. Similarity and Diversity Induced Paired Projection for Cross-Modal Retrieval. *Information Sciences* 539 (2020), 215–228.

[16] Shuang Li, Tong Xiao, Hongsheng Li, Wei Yang, and Xiaogang Wang. 2017. Identity-aware textual-visual matching with latent co-attention. In *Proceedings of the IEEE International Conference on Computer Vision*. 1890–1899.

[17] Shuang Li, Tong Xiao, Hongsheng Li, Bolei Zhou, Dayu Yue, and Xiaogang Wang. 2017. Person Search with Natural Language Description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1970–1979.

[18] Kaiyi Lin, Xing Xu, Lianli Gao, Zheng Wang, and Heng Tao Shen. 2020. Learning Cross-Aligned Latent Embeddings for Zero-Shot Cross-Modal Retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 11515–11522.

[19] Jiawei Liu, Zheng-Jun Zha, Richang Hong, Meng Wang, and Yongdong Zhang. 2019. Deep Adversarial Graph Attention Convolution Network for Text-Based Person Search. In *Proceedings of the 27th ACM International Conference on Multimedia*. 665–673.

[20] Asad Munir, Chengjin Lyu, Bart Goossens, Wilfried Philips, and Christian Micheloni. 2021. Resolution based Feature Distillation for Cross Resolution Person Re-Identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 281–289.

[21] Lu Pang, Yaowei Wang, Yi-Zhe Song, Tiejun Huang, and Yonghong Tian. 2018. Cross-Domain Adversarial Feature Learning for Sketch Re-Identification. In *Proceedings of the 26th ACM International Conference on Multimedia*. 609–617.

[22] Yuxin Peng and Jinwei Qi. 2019. CM-GANs: Cross-Modal Generative Adversarial Networks for Common Representation Learning. *ACM Transactions on Multimedia Computing, Communications, and Applications* 15, 1 (2019), 1–24.

[23] Haopeng Qiang, Yuan Wan, Lun Xiang, and Xiaojing Meng. 2020. Deep Semantic Similarity Adversarial Hashing for Cross-Modal Retrieval. *Neurocomputing* 400 (2020), 24–33.

[24] Amaia Salvador, Nicholas Hynes, Yusuf Aytar, Javier Marin, Ferda Ofli, Ingmar Weber, and Antonio Torralba. 2017. Learning Cross-Modal Embeddings for Cooking Recipes and Food Images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3020–3028.

[25] Jingyu Shao, Qing Wang, Asiri Wijesinghe, and Erhard Rahm. 2020. ErGAN: Generative Adversarial Networks for Entity Resolution. In *2020 IEEE International Conference on Data Mining*. IEEE, 1250–1255.

[26] Jifei Song, Yi-Zhe Song, Tony Xiang, and Timothy M Hospedales. 2017. Fine-Grained Image Retrieval: the Text/Sketch Input Dilemma. In *BMVC*, Vol. 1. 2.

[27] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and Tell: A Neural Image Caption Generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3156–3164.

[28] Bokun Wang, Yang Yang, Xing Xu, Alan Hanjalic, and Heng Tao Shen. 2017. Adversarial Cross-Modal Retrieval. In *Proceedings of the 25th ACM International Conference on Multimedia*. 154–162.

[29] Guan-An Wang, Tianzhu Zhang, Yang Yang, Jian Cheng, Jianlong Chang, Xu Liang, and Zeng-Guang Hou. 2020. Cross-Modality Paired-Images Generation for RGB-Infrared Person Re-Identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 12144–12151.

[30] Tan Wang, Xing Xu, Yang Yang, Alan Hanjalic, Heng Tao Shen, and Jingkuan Song. 2019. Matching Images and Text with Multi-Modal Tensor Fusion and Re-Ranking. In *Proceedings of the 27th ACM International Conference on Multimedia*. 12–20.

[31] Zheng Wang, Zhixiang Wang, Yinqiang Zheng, Yang Wu, Wenjun Zeng, and Shin'ichi Satoh. 2019. Beyond Intra-Modality: A Survey of Heterogeneous Person Re-Identification. *arXiv preprint arXiv:1905.10048* (2019).

[32] Fei Wu, Xiao-Yuan Jing, Zhiyong Wu, Yimu Ji, Xiwei Dong, Xiaokai Luo, Qinghua Huang, and Ruchuan Wang. 2020. Modality-Specific and Shared Generative Adversarial Network for Cross-Modal Retrieval. *Pattern Recognition* 104 (2020), 107335.

[33] Jun Xiao, Hao Ye, Xiangnan He, Hanwang Zhang, Fei Wu, and Tat-Seng Chua. 2017. Attentional Factorization Machines: Learning the Weight of Feature Interactions via Attention Networks. *arXiv preprint arXiv:1708.04617* (2017).

[34] Xing Xu, Jingkuan Song, Huimin Lu, Yang Yang, Fumin Shen, and Zi Huang. 2018. Modal-Adversarial Semantic Learning Network for Extendable Cross-Modal Retrieval. In *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*. 46–54.

[35] Lan Yan, Wenbo Zheng, Fei-Yue Wang, and Chao Gou. 2021. Weakly Supervised Sketch Based Person Search. In *Proceedings of the 2021 International Conference on Multimedia Retrieval*. 491–495.

[36] Mang Ye, Chao Liang, Zheng Wang, Qingming Leng, Jun Chen, and Jun Liu. 2015. Specific Person Retrieval via Incomplete Text Description. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*. 547–550.

[37] Zhou Yin, Wei-Shi Zheng, Ancong Wu, Hong-Xing Yu, Hai Wan, Xiaowei Guo, Feiyue Huang, and Jianhuang Lai. 2017. Adversarial Attribute-Image Person Re-Identification. *arXiv preprint arXiv:1712.01493* (2017).

[38] Qian Yu, Feng Liu, Yi-Zhe Song, Tao Xiang, Timothy M Hospedales, and Chen-Change Loy. 2016. Sketch Me That Shoe. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 799–807.

[39] Yawen Zeng, Da Cao, Shaofei Lu, Hanling Zhang, Jiao Xu, and Qin Zheng. 2022. Moment is Important: Language-Based Video Moment Retrieval via Adversarial Learning. *ACM Trans. Multim. Comput. Commun. Appl.* 18 (2022), 56:1–56:21.

[40] Yawen Zeng, Da Cao, Xiaochi Wei, Meng Liu, Zhou Zhao, and Zheng Qin. 2021. Multi-Modal Relational Graph for Cross-Modal Video Moment Retrieval. In *CVPR*. IEEE, 2215–2224.

[41] Ying Zhang and Huchuan Lu. 2018. Deep Cross-Modal Projection Learning for Image-Text Matching. In *Proceedings of the European Conference on Computer Vision*. 686–701.

[42] Kecheng Zheng, Wu Liu, Lingxiao He, Tao Mei, Jiebo Luo, and Zheng-Jun Zha. 2021. Group-aware label transfer for domain adaptive person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5310–5319.

[43] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. 2015. Scalable Person Re-Identification: A Benchmark. In *Proceedings of the IEEE International Conference on Computer Vision*. 1116–1124.

[44] Zhedong Zheng, Xiaodong Yang, Zhiding Yu, Liang Zheng, Yi Yang, and Jan Kautz. 2019. Joint Discriminative and Generative Learning for Person Re-Identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2138–2147.

[45] Aichun Zhu, Zijie Wang, Yifeng Li, Xili Wan, Jing Jin, Tian Wang, Fangqiang Hu, and Gang Hua. 2021. DSSL: Deep Surroundings-person Separation Learning for Text-based Person Retrieval. In *Proceedings of the 29th ACM International Conference on Multimedia*. 209–217.

[46] Bin Zhu, Chong-Wah Ngo, Jingjing Chen, and Yanbin Hao. 2019. R2gan: Cross-modal Recipe Retrieval with Generative Adversarial Network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11477–11486.

[47] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. In *Proceedings of the IEEE International Conference on Computer Vision*. 2223–2232.