This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Towards Modality-Agnostic Person Re-identification with Descriptive Query

Cuiqun Chen[1], Mang Ye[1,2]*, Ding Jiang[1]
[1]National Engineering Research Center for Multimedia Software, Institute of Artificial Intelligence,
Hubei Key Laboratory of Multimedia and Network Communication Engineering,
School of Computer Science, Wuhan University, Wuhan, China
[2] Hubei Luojia Laboratory, Wuhan, China

https://github.com/ccq195/UNIReID

Abstract

Person re-identification (ReID) with descriptive query (text or sketch) provides an important supplement for general image-image paradigms, which is usually studied in a single cross-modality matching manner, e.g., text-to-image or sketch-to-photo. However, without a camera-captured photo query, it is uncertain whether the text or sketch is available or not in practical scenarios. This motivates us to study a new and challenging modality-agnostic person re-identification problem. Towards this goal, we propose a unified person re-identification (UNIReID) architecture that can effectively adapt to cross-modality and multi-modality tasks. Specifically, UNIReID incorporates a simple dual-encoder with task-specific modality learning to mine and fuse visual and textual modality information. To deal with the imbalanced training problem of different tasks in UNIReID, we propose a task-aware dynamic training strategy in terms of task difficulty, adaptively adjusting the training focus. Besides, we construct three multi-modal ReID datasets by collecting the corresponding sketches from photos to support this challenging study. The experimental results on three multi-modal ReID datasets show that our UNIReID greatly improves the retrieval accuracy and generalization ability on different tasks and unseen scenarios.

1. Introduction

Person re-identification [42] involves using computer vision techniques to identify pedestrians in video and still images. Given a monitored pedestrian image/video or a text description, ReID aims to retrieve all images/videos of that pedestrian across devices. ReID is widely used in intelligent video surveillance, intelligent security, and other fields. The existing ReID researches include single-modal ReID [4, 13, 21, 47, 48] and cross-modal ReID [2, 17, 41]. The former is restricted to retrieval between RGB images, whereas the latter allows retrieval of RGB images based on different query modalities (*e.g.*, IR, text, or sketch). Particularly, the appearance of a suspect may only be described verbally in many criminal cases. Person re-identification with descriptive query (text or sketch) is well suited for these scenarios and has greater research value for grounded applications of ReID models.

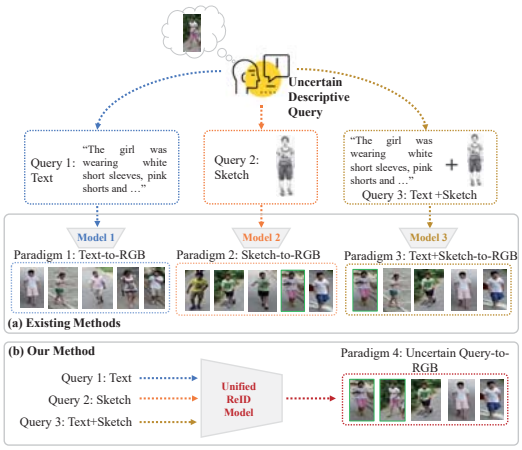

Figure 1. Illustration of our idea. Existing ReID methods [17, 24, 43] yield different paradigm models for different descriptive queries. However, it is uncertain whether the text or sketch is available or not in practical scenarios. Our unified ReID model enables target pedestrian search under uncertain query inputs. The green boxes match the query.

Most existing text-based or sketch-based person re-identification methods rely on only one of the modalities as a query set to achieve pedestrian retrieval. Although the text modality is relatively easy to access, it fails to accurately depict visual appearance details, *i.e.*, coarse-grained representations [29]. As the saying goes, a picture is worth a thousand words, and the sketch image of a pedestrian is closer to the visual modality, showing specific information about

*Corresponding Author: Mang Ye (yemang@whu.edu.cn)

the target pedestrian, such as structural information. Since each task is trained independently, as shown in Fig 1(a), it is impossible to generalize to unseen modalities. For example, a ReID model trained on text-based datasets is basically invalid on sketch-based scenes, and vice versa. This greatly limits the applicability of existing methods for practical model deployment.

Meanwhile, multi-modal fusion has been proven to be an important technique to improve model accuracy in computer vision, *e.g.*, multi-modal face anti-spoofing [11], multi-modal generation and understanding [18]. Recently, Zhai *et al.* [43] first propose to implement multi-modal ReID using both sketch and text modalities as the query. Their experimental results indicate that the combination of text and sketch modalities enhances the performance of the ReID model. However, this method adopts independent text and image pre-training parameters for multi-modal representation learning, which has poor generalizability and yields low accuracy. More importantly, it is often uncertain whether text or sketch is available in a real scenario, *i.e.*, modality deficit problems often arise when a specific modality is not available. Due to the independent training of tasks, existing cross-modal or multi-modal ReID methods are difficult to handle this problem. A smarter surveillance system should be capable of handling various modalities of information efficiently. Therefore, in this paper, we propose the concept of modality-agnostic person re-identification to handle the modality uncertainty issue.

Specifically, we design a unified person re-identification architecture (UNIReID) in Fig 1(b), which is effective in both cross-modal and multi-modal applications. The greatest challenge in unifying learning across modalities is to mine a shared semantic representation space. At first, we propose a task-specific modality learning scheme to support individual task learning. Essentially, this scheme considers unified person re-identification as a set of retrieval tasks involving Text-to-RGB, Sketch-to-RGB, and Text+Sketch-to-RGB. Inspired by CLIP [25] which is a pre-trained model for matching image and text modalities, UNIReID uses a simple dual-encoder for visual modality and textual modality feature extraction and fusion. All visual modalities share a single image encoder. For multi-modal ReID, we fuse the sketch and text modalities into a single query by a simple feature-level summation. Task-specific metric learning explicitly minimizes the feature distances between various types of query samples and gallery samples to learn modality-shared feature representations.

In addition, considering that tasks have varying difficulties, unified ReID faces an additional challenge in balancing learning among tasks, which may result in the overfitting of individual tasks. To handle this problem, we design a task-aware dynamic training strategy that adaptively adjusts for training imbalances between tasks. The rationale for dynamic training is to modulate the loss contribution of different retrieval tasks according to the training difficulty of tasks (*i.e.*, prediction confidence). Our dynamic training strategy improves generalization capability by tending to train difficult tasks. Finally, a cross-modality interaction is designed to align sketch and text feature representations. In view of the differences between sketch and text modal features, we minimize the similarity distribution distances between sketch-RGB and text-RGB pairs to align modality information for modality fusion. With the help of rich multi-modal data, our model achieves mutual enhancement of tasks and improves the robustness against diverse query modality variations.

To facilitate the modality-agnostic ReID study, we construct three multi-modal ReID datasets. Concretely, based on the text-based ReID datasets, namely CUHK-PEDES [17], ICFG-PEDES [7], and RSTPReid [50], we collect the sketch modality for each identity. Considering that sketching by hand is time-consuming and labor-intensive, we propose to generate the sketch modality for each identity from the photo modality. The detailed collection information for the datasets can be found in Section 3.

The main contributions of this paper are listed below:

- We start the first attempt to investigate the modality-agnostic person re-identification with the descriptive query, which provides a flexible solution to deal with query uncertainty in real-world scenarios.
- We introduce a novel unified person re-identification (UNIReID) architecture based on a dual-encoder to jointly integrate cross-modal and multi-modal task learning. With task-specific modality learning and task-aware dynamic training, UNIReID enhances generalization ability across tasks and domains.
- We contribute three multi-modal ReID datasets to support unified ReID evaluation. Extensive experiments on both multi-modal matching and generalized cross-modality matching have verified the advantage of UNIReID, achieving much higher accuracy than existing counterparts.

2. Related Work

2.1. Cross-modal Person Re-identification

Person re-identification (ReID) aims to retrieve all images of a target pedestrian across devices and emphasizes learning discriminative pedestrian representations. According to the different modalities of pedestrian information representation, person re-identification can be classified into single-modal ReID [4,42] and cross-modal ReID [2,41,50]. Specifically, cross-modal ReID considers some special scenarios in which RGB images of pedestrians are not immediately available and proposes using non-RGB modalities

(such as infrared images [3,38,40,41], text [7,10,17,29,50], and sketch [2,12,24,39]) to characterize pedestrian information, enlarging the application scenario of ReID technology.

In many criminal cases, the staff usually searches the target pedestrian image directly based on the natural language descriptions (text) of the witness, or indirectly by drawing a sketch of the pedestrian based on the textual descriptions. Li *et al.* [17] first propose to explore the problem of retrieving the target pedestrians with natural language descriptions for adaptation to real-world circumstances. Several methods [7,10] propose aligning the feature representation of image and text modalities by using attention mechanisms. Recently, Shao *et al.* [29] analyze the granularity differences between the visual modality and textual modality and propose a granularity-unified representation learning method for text-based ReID.

On the other hand, Pang *et al.* [24] first propose to use professional sketches as queries to search for the target person in the RGB gallery. They design cross-domain adversarial learning methods to mine domain-invariant feature representations. Chen *et al.* [2] present a novel asymmetrical disentanglement scheme that resolves the information asymmetry issue between sketch and photo modalities and enhances the performance of the model. In order to explore the complementarity between the sketch modality and the text modality, Zhai *et al.* [43] introduce a multi-modal ReID task that combines both sketch and text modalities as queries for retrieval. However, existing methods for different descriptive queries are independent learning without considering the query uncertainty, leading to poor generalizability of models in real scenarios. Meanwhile, they fail to explore the mutual enhancement between tasks and have limited discriminatory power. In this paper, we propose a unified person re-identification architecture for handling cross-modal and multi-modal retrieval tasks together.

2.2. Vision-Language Models

Recently, natural language processing (NLP) and computer vision have both benefited greatly from transformer models. Many vision-language pre-training (VLP) researches, such as image-text, apply the transformer as the model architecture. The vision-language pre-training methods can be classified from a multi-modal fusion perspective as single-stream methods and dual-stream methods. The single-stream methods [5, 6, 16] use cascading text and image results as input to the network. Conversely, the dual-stream methods [9,20,25,31,44] transmit the visual and textual modalities into different networks, respectively. Lu *et al.* [20] claim that vision and text modalities should require different encoding depths, and design a dual-stream feature extraction framework. Radford *et al.* [9] employ text information to supervise the visual task self-training, thereby essentially changing the classification task into an image-text matching task and significantly enhancing model performance. Encouraged by this method, to handle the multi-modal data in a unified person re-identification framework, we adopt a dual-encoder based on a transformer for feature learning of visual and textual modalities.

Figure 2. Illustration of the generated sketches. All images are selected from Tri-CUHK-PEDES and Tri-ICFG-PEDES datasets.

3. Multi-modal Dataset Construction

Considering that there is no publicly available large-scale multi-modal person re-identification dataset, we first construct three multi-modal datasets, *i.e.*, Tri-CUHK-PEDES, Tri-ICFG-PEDES, and Tri-RSTPReid, to facilitate the community research. These three datasets are extensions of text-based datasets CUHK-PEDES [17], ICFG-PEDES [7], and RSTPReid [50], respectively. Since the process of collecting professional sketch images is time-consuming and expensive, Zhai *et al.* [43] use an active learning approach to generate text descriptions corresponding to RGB images based on sketch ReID dataset [24]. However, this may raise two problems: 1) It is challenging to generate accurate and comprehensive pedestrian descriptions from images due to the complex process of converting images into text. 2) Since the sketch person re-identification dataset is small, it is easy to overfit the training set and result in poor generalization.

Different from this, our proposal is to construct multi-modal datasets by expanding the text datasets with the sketch modality. Comparatively to the sketch dataset, the text dataset typically contains a greater amount of pedestrian data, which is advantageous for generalizability studies. More importantly, it is significantly easier to generate a sketch from a photograph than to create a text description, and there are a number of mature tools and APIs available.

The generated sketch includes a more comprehensive representation of pedestrian information. Thus, we propose to obtain the sketch modality information of the corresponding pedestrians by the following two steps:

1) **Background Erasing**. With reference to the hand sketch, we first erase the background of RGB images through the Aliyun API[1] to mitigate the impact of the background noise. This API returns the foreground pedestrian image by identifying the human silhouette in the input image and separating it from the background.

2) **Sketch Synthesis**. In recent years, there has been extensive research on sketch synthesis, one of the capabilities humans hope machines will emulate. Based on our investigations, we find that Meitu API[2] is more suitable for pedestrian images of lower resolution. On the basis of the RGB image generated by removing the background in step one, we apply the Meitu API to generate hand-drawn sketch-style sketches of the pedestrian. As shown in Fig 2, we present some generated sketch examples. The statistics of the proposed three datasets are shown in Table 1.

Table 1. **Dataset statistics**.

Datasets	#ID	#RGB	#Text	#Sketch
Tri-CUHK-PEDES	13003	40206	80440	40206
Tri-ICFG-PEDES	4102	54522	54522	54522
Tri-RSTPReid	4101	20505	41010	20505

4. Unified Person Re-identification

4.1. Overall Architecture

The key to achieving the unified person search with different descriptive queries is to build a shared feature extractor for multi-modal data. Given a descriptive query sample, it is desirable that the feature extractor could adaptively learn the modality-invariant information for each modality to retrieve the target pedestrians. Generally, pedestrian descriptions that include RGB, sketch, or text can be categorized as either vision or language modality. Thus, the unified person re-identification is primarily concerned with modeling the relationships between and within both modalities. In recent years, several vision-language pre-training models [6, 9, 16, 25] have been proposed for learning the semantic correspondence between image and text modalities. Specifically, CLIP [25] model utilizes text as a supervised signal for learning text-image matching relationships, resulting in a transferrable model. Based on this model, we propose a dual-encoder transformer architecture for multi-modality feature learning. As shown in Fig 3, all visual modalities share the same encoder parameters.

For a visual input V (RGB V_r or Sketch V_s), it is first partitioned into equal-sized patches, and then each patch is mapped to a vector with fixed dimensions, obtaining the image sequence embedding $V = \{[IMG], v_1, v_2, ..., v_m\}$. After combining with the positional embedding, it is fed into the vision transformer [8] to learn visual contextual feature representations. The $V[IMG]$ represents the global feature representation of an image. Similarly, for a textual description T, it is first encoded to a word sequence $T = \{[SOS], t_1, t_2, ..., t_n, [CLS]\}$ using the Byte-Pair Encoding method [27]. A text transformer encoder [33] is adopted to mine textual contextual relations. The $T[SOS]$ and $T[CLS]$ tokens are used to indicate the beginning and end of the textual sequence feature, respectively. Finally, the global semantic feature tokens $V_r[IMG]$, $V_s[IMG]$, and $T[CLS]$ are used as the modality representations for cross-modality matching.

Moreover, benefiting from the unified image and text pre-training parameters, we combine information from both sketch and text modalities using a simple summation-fusion strategy. There is a theoretical possibility that uniform modal parameters could mitigate the effects of modal differences during training. Due to the lengths of visual and textual tokens being different, we apply the global tokens for information fusion. Thus, the final fusion feature can be defined as $F[CLS] = V_s[IMG] + T[CLS]$. Compared with the descriptive semantic fusion method [43], our fusion method demonstrates the advantages of a straightforward implementation and low computational complexity.

4.2. Task-specific Modality Learning

Towards unified cross-modal learning, it is essential to mine the modality-invariant representations among the three modalities. It is possible to reach this goal by training three types of retrieval tasks together (*i.e.*, Sketch-to-RGB, Text-to-RGB, and Text+Sketch-to-RGB). A few cross-modal matching methods [2, 41, 43] generally align the cross-modality representations under the guidance of label information. Aware of the migration nature of the model, some others [25, 45] employ a contrastive representation learning scheme to explore a cross-modal embedding space. These methods create a large number of sample pairs across modalities and maximize the cosine similarity between true sample pairs and minimize the cosine similarity between incorrect sample pairs. This contrastive optimization enables the feature encoder to preserve as much mutual information as possible between the true pairs.

Based on this idea, we propose a task-specific modality learning scheme to independently optimize the distances between different query samples and gallery samples. With specific task learning, our unified model allows testing to account for missing modalities. Specifically, given the query modality feature $\mathbf{q}$ and gallery modality feature $\mathbf{g}$, we sample M pairs (query, gallery) in a batch. Due to the modality variations, the contrastive training function in the

[1] https://vision.aliyun.com/
[2] https://ai.meitu.com/

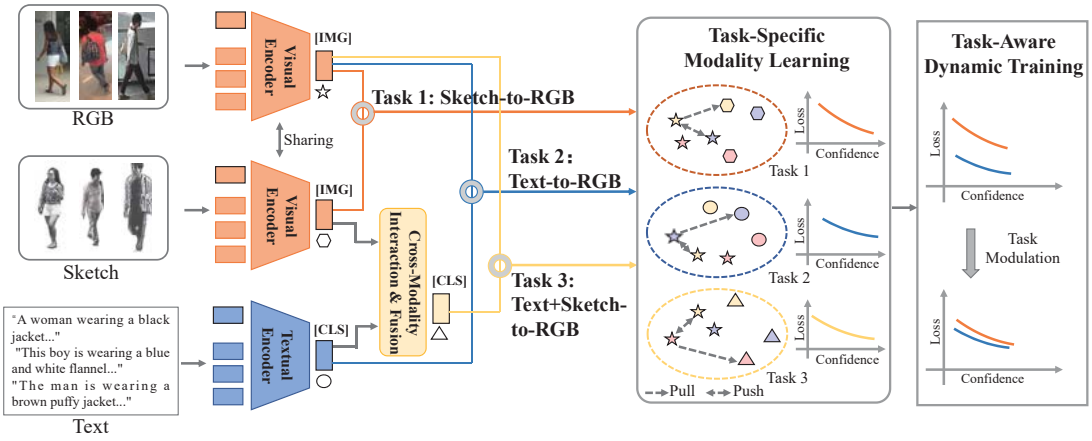

Figure 3. The network architecture of the proposed UNIReID. All visual modalities share a single visual encoder. With task-specific modality learning and task-aware dynamic training, UNIReID facilitates task mutuality and improves model generalization.

same form as the infoNCE loss [23] is asymmetrical and consists of the average of two components

$$\mathcal{L}^{(q\to g)}(i) = -\log \frac{\exp\left(\langle \mathbf{q}_i, \mathbf{g}_i \rangle / \tau\right)}{\sum_{k=1}^{M} \exp\left(\langle \mathbf{q}_i, \mathbf{g}_k \rangle / \tau\right)}, \tag{1}$$

$$\mathcal{L}^{(g\to q)}(i) = -\log \frac{\exp\left(\langle \mathbf{g}_i, \mathbf{q}_i \rangle / \tau\right)}{\sum_{k=1}^{M} \exp\left(\langle \mathbf{g}_i, \mathbf{q}_k \rangle / \tau\right)}, \tag{2}$$

where $(\mathbf{q}_i, \mathbf{g}_i)$ represents the i-th pair. $\langle \mathbf{q}_i, \mathbf{g}_i \rangle$ defines the cosine similarity between query and gallery features. τ is the temperature parameter. In our unified model, we consider three types of retrieval tasks depending on the descriptive query modality, *i.e.*, Sketch-to-RGB ($\mathcal{L}_{S\to R}$), Text-to-RGB ($\mathcal{L}_{T\to R}$), and Text+Sketch-to-RGB ($\mathcal{L}_{F\to R}$). Finally, the task-specific modality loss ($\mathcal{L}_s$) can be formulated as

$$\begin{aligned}
\mathcal{L}_s &= \mathcal{L}_{S\to R} + \mathcal{L}_{T\to R} + \mathcal{L}_{F\to R} \\
&= \frac{1}{M}\sum_{i=1}^{M} \frac{1}{2}\mathcal{L}^{(V_s[IMG]\to V_r[IMG])}(i) + \frac{1}{2}\mathcal{L}^{(V_r[IMG]\to V_s[IMG])}(i) \\
&+ \frac{1}{M}\sum_{i=1}^{M} \frac{1}{2}\mathcal{L}^{(T[CLS]\to V_r[IMG])}(i) + \frac{1}{2}\mathcal{L}^{(V_r[IMG]\to T[CLS])}(i) \\
&+ \frac{1}{M}\sum_{i=1}^{M} \frac{1}{2}\mathcal{L}^{(F[CLS]\to V_r[IMG])}(i) + \frac{1}{2}\mathcal{L}^{(V_r[IMG]\to F[CLS])}(i).
\end{aligned} \tag{3}$$

Although individual task mines the discriminative modality-specific information, using equal task weighting factors may not be an ideal solution for unified model learning. This issue is discussed in more detail below. Considering that the difference between each query modality and the gallery modality is different, the difficulty of optimizing for different search tasks varies. It is possible to train some tasks incompletely and overfit some tasks using equal loss weights. Meanwhile, some prior efforts on multi-task learning [28, 32] combine multiple target tasks by using a weighted linear sum of losses. However, a significant amount of time and effort must be spent on tuning the model since it is highly sensitive to the selection of weight parameters. Therefore, it is desirable to learn more robust multi-modal features by means of a more convenient method. After that, we describe how to dynamically determine the optimal multi-task weights by considering the predicted confidence of different query modalities.

4.3. Task-aware Dynamic Training

The rationale of our unified ReID model is to learn the discriminative feature representations of visual and textual modalities. Through joint learning across multiple tasks, we intend to achieve task mutuality, increasing the generalization ability of the model in different modalities. However, it is possible that a simple fusion approach with fixed weights may not provide adequate training for some cross-modal retrieval tasks, leading to limited robustness.

To address this issue, we propose a task-aware dynamic training method to balance multi-task training. Essentially, the idea is that when one cross-modal retrieval task is able to retrieve corresponding images or texts with high predicted confidence, the other cross-modal retrieval task can increase its contribution to the loss. In our unified ReID model, the Sketch-to-RGB and Text-to-RGB retrieval tasks are two fundamental tasks that directly guide the learning of robust feature representations in both visual and textual modalities. Thus, the task-aware dynamic training forces the performance improvement of these two retrieval tasks, which can then be utilized with the multi-modal retrieval. It is well known that the Focal Loss [19] reduces the loss contribution of correctly classified samples by adding a modulation factor to adjust for data imbalances. Specifically, the modulation factor resulting from the predicted probability of the sample indicates the degree of difficulty of sample classification. Following this idea, we modulate the loss contribu-

tions according to the prediction confidence of tasks.

Considering that the cosine similarity after a softmax function between the query and gallery samples in **eq.**1 determines the prediction confidence of a true match. With the task-specific modality loss, we could obtain the prediction confidences for the Sketch-to-RGB and Text-to-RGB retrieval tasks, as follows

$$p_{SR}(i) = \exp(-\mathcal{L}_{S\to R}(i)), \tag{4}$$

$$p_{TR}(i) = \exp(-\mathcal{L}_{T\to R}(i)). \tag{5}$$

In the case of a task that provides high prediction confidence, this enhances the loss of contribution of the other task. Referring to [11], we calculate the modulation factors by taking the harmonic mean of both two tasks and multiplying it by the predicted confidence of the other task. And the modulation factor formulation can be defined by

$$w_{SR}(i) = p_{TR}(i) * \frac{2 * p_{SR}(i) * p_{TR}(i)}{p_{SR}(i) + p_{TR}(i)}, \tag{6}$$

$$w_{TR}(i) = p_{SR}(i) * \frac{2 * p_{SR}(i) * p_{TR}(i)}{p_{SR}(i) + p_{TR}(i)}. \tag{7}$$

With the task modulation factors, the task-specific modality loss could be updated by

$$\mathcal{L}_{S\to R}(i) = \alpha_{\mathrm{t}} \left(1 + w_{SR}(i)\right)^{\gamma} \mathcal{L}_{S\to R}(i), \tag{8}$$

$$\mathcal{L}_{T\to R}(i) = \alpha_{\mathrm{t}} \left(1 + w_{TR}(i)\right)^{\gamma} \mathcal{L}_{T\to R}(i), \tag{9}$$

where α_{t} and γ are the hyper-parameters to control the decay trend of the loss curve. Throughout all of our experiments, we employ the empirical values of $\alpha_{\mathrm{t}} = 1$ and $\gamma = 3.5$. When the predicted confidence of one task is zero (*e.g.*, $w_{SR}(i) \to 0$), it indicates that the other task is inadequate training and that the loss is a standard infoNCE loss. Conversely, the loss contributions will increase at a larger rate as w_{SR} or w_{TR} increases, aiming at achieving the multi-task generalization.

4.4. Cross-modality Interaction

For the multi-modal fusion task, we further design a cross-modality interaction to align the feature representations before modality information fusion. More formally, when a sample has a pair of sketch and text features that provide complementary information to each other, it is possible to improve model retrieval accuracy by combining two modality features. Currently, available fusion methods [11, 43] do not consider the interaction between the two modalities prior to fusion, but only consider how the fusion features are derived. It would be difficult to mine cross-modal shared information, limiting the discriminating power of fusion features.

To avoid this, our cross-modality interaction proposes to align the feature distributions between sketch and text modalities. Considering the large visual and textual modality discrepancy, the explicit feature alignment may result in the loss of modality-specific information. In this paper, we align similarity distribution between Sketch-RGB and Text-RGB pairs to mine modality-shared information for feature fusion. The basic idea is that for different query modality inputs (sketch or text), the similarity distributions between them and the gallery modality (RGB) should be the same. The cross-modality interaction loss $\mathcal{L}_c$ can be denoted as

$$\mathcal{L}_c = -\frac{1}{M}\sum\nolimits_{i=1}^{M} f(P_{TR}(i))\log(f(Q_{SR}(i))), \tag{10}$$

where Q_{SR} and P_{TR} denote the cosine similarity distributions between Sketch-RGB and Text-RGB feature pairs in a batch, respectively. $f(\cdot)$ represents the softmax function used to map inputs to probabilities in the range [0,1]. This loss maintains the same level of prediction confidence across queries.

5. Experiments

5.1. Experimental Settings

Datasets. We evaluate our UNIReID on the proposed three multi-modal datasets, including Tri-CUHK-PEDES, Tri-ICFG-PEDES, and Tri-RSTPReid. An overview of training and test set partitioning for each dataset can be found in the existing work [7, 17, 50].

Evaluation Protocols. Following existing cross-modality ReID settings [2, 41, 42], we use the Rank-k matching accuracy, mean Average Precision (mAP), and mean Inverse Negative Penalty (mINP) [42] metrics for performance evaluation in our UNIReID.

Implementation Details. We employ the Vision Transformer [8] as the visual modalities feature learning backbone, and the Transformer model [33] as the textual modality feature learning backbone. Both backbones have pretrained parameters derived from CLIP [25]. In a batch, we randomly select 64 identities, each containing a sketch, a text, and an RGB sample. Also, we apply random horizontal flipping, padding, random cropping, and random erasing as data augmentation for visual modality. We train our UNIReID model with the Adam optimizer for 60 epochs. And the initial learning rate is computed as 1e-5 and decayed by a cosine schedule. The temperature parameter τ is set as 0.07. All experiments are supported by Huawei MindSpore [1].

5.2. Ablation Study

We conduct extensive experiments to evaluate each element of our method. For each experiment/model, we test three retrieval tasks, *i.e.*, Text-to-RGB (T→R), Sketch-to-RGB (S→R), and Sketch+Text-to-RGB (T+S→R). Additionally, the RGB images used for generating the sketches are removed from the gallery for three datasets during the

Table 2. Ablation study about each component on three multi-modal datasets. Rank (R) at k accuracy (%), mAP(%), and mINP(%) are reported.

Tasks	Methods	Tri-CUHK-PEDES			Tri-ICFG-PEDES			Tri-RSTPReid		
		R1	mAP	mINP	R1	mAP	mINP	R1	mAP	mINP
T→R	$\mathcal{L}_{T\to R}$	52.17	51.35	41.81	52.09	31.06	5.41	47.60	40.51	23.85
	$\mathcal{L}_s$	51.06	50.73	41.41	50.68	29.54	5.01	47.55	39.47	22.34
	w Dynamic	53.48	53.01	43.60	55.04	33.06	6.13	49.15	41.53	24.59
	w $\mathcal{L}_c$	53.82	53.43	44.28	55.39	33.79	6.27	49.30	41.67	24.69
S→R	$\mathcal{L}_{S\to R}$	58.18	44.85	28.09	46.49	1.41	0.20	31.10	17.58	4.12
	$\mathcal{L}_s$	80.70	72.36	59.29	70.11	29.48	2.82	60.10	44.10	20.80
	w Dynamic	84.02	76.79	65.63	76.15	37.73	6.05	64.90	50.77	27.40
	w $\mathcal{L}_c$	84.87	78.85	68.55	77.47	40.41	6.31	65.80	51.22	27.47
T+S→R	$\mathcal{L}_{F\to R}$	63.94	51.14	34.04	38.00	22.35	4.98	53.86	13.21	0.45
	$\mathcal{L}_s$	85.41	78.45	67.23	78.41	38.90	5.31	69.80	53.52	28.88
	w Dynamic	86.14	80.20	70.17	81.96	44.91	8.55	73.05	58.42	34.38
	w $\mathcal{L}_c$	86.29	80.92	71.30	82.17	47.00	8.74	73.20	58.72	34.61

inference phase to avoid over-similarity in the sketch and photo information.

Independent Individual Tasks Learning (IITL). With the specific visual and textual encoder, we first train the individual tasks independently through the infoNCE loss, *i.e.,* $\mathcal{L}_{T\to R}$, $\mathcal{L}_{S\to R}$, *and* $\mathcal{L}_{F\to R}$. From the results in Table 2, we find that the dual-encoder model is appropriate for three retrieval tasks and multi-modal fusion could effectively improve the accuracy of the model.

Effect of Task-specific Modality Learning. Considering the uncertainty of whether the text or sketch is available in practical scenarios, we design a unified ReID model for multi-modality information fusion and multi-task learning. Through the task-specific modality learning ($\mathcal{L}_s$), our unified ReID model obtains 85.41%, 78.41%, and 69.80% Rank-1 accuracy with the Text+Sketch query on three datasets, respectively. In contrast, the simple task-based loss fusion does not result in significant performance improvements in text-based retrieval.

Effect of Task-aware Dynamic Training. To keep the task training balanced for multi-task learning, we propose a task-aware dynamic training strategy (w Dynamic) that modulates the loss contributions of different tasks. From the results in Table 2, we find that the dynamic training strategy outperforms the fixed weights optimization ($\mathcal{L}_s$) by a large margin. For example, this strategy achieves 1.31%, 2.95%, and 1.55% improvements in the Rank-1 accuracy of the Text-to-RGB task on three datasets, respectively. In addition, compared with independent individual task learning, our multi-task joint learning with task-aware dynamic training brings in the improvement of retrieval accuracy, verifying the generalization ability.

Effect of Cross-modality Interaction. For multi-modal fusion, we perform the cross-modality interaction (w $\mathcal{L}_c$) before modality information fusion to facilitate modality-shared feature learning. As shown in Table 2, it is clear that the cross-modality interaction further improves retrieval accuracy for three tasks under various metrics. It obtains 44.28%, 68.55%, and 71.30% mINP for the three retrieval tasks on the Tri-CUHK-PEDES dataset, respectively.

5.3. Comparison with State-of-the-art Methods

As shown in Tables 3, 4, 5, we compare our method with the text-based state-of-the-art methods. For a fair comparison with existing methods, we test our IITL and UNIReID models on Text-to-RGB retrieval tasks by using the original gallery set of three datasets (IITL (T→R)* and UNIReID (T→R)*). From these results, the following two points may be summarized: 1) The proposed dual-encoder for visual and textual modalities feature learning is more effective than most CNN-based methods. 2) Our UNIReID with a dynamic training strategy considers cross-modality tasks as well as multi-modality tasks in a unified framework, resulting in complementarity between tasks and improvements in retrieval performance.

5.4. Cross-domain Generalization Evaluation

The generalization ability between different domains is a key measure of the ReID model, which is significant for practical surveillance systems. Existing methods employ the cross-dataset evaluation [4, 29] to validate the generalization ability of the model. In this paper, we calculate the retrieval accuracy on the PKU-Sketch dataset [24] using UNIReID model parameters from the Tri-CUHK-PEDES dataset. PKU-Sketch is the only dataset that includes hand-drawn professional pedestrian sketches. Zhai *et al.* [43] collect the textual description for each identity through an image captioning model. Similar to [2], we perform ten tests and analyze the average results to avoid randomness in the evaluation of performance.

As shown in Table 6, we first compare our UNIReID with some existing sketch-based methods, such as CD-

Table 3. Comparison with the state-of-the-arts on Tri-CUHK-PEDES dataset. Rank (R) at k accuracy (%) is reported. $*$ indicates that the original gallery set is used for testing.

Methods	Venue	R1	R5	R10
CMPM/C [46]	ECCV18	49.37	-	79.27
TIMAM [26]	ICCV19	54.51	77.56	84.78
GLAM [14]	AAAI20	54.12	75.45	82.97
ViTAA [35]	ECCV20	55.97	75.84	83.52
MGEL [34]	IJCAL21	60.27	80.01	86.74
DSSL [50]	MM21	59.98	80.41	87.56
IVT [30]	Arxiv22	65.59	83.11	89.21
LBUL+BERT [37]	MM22	64.04	82.66	87.22
CAIBC [36]	MM22	64.43	82.87	87.35
LGUR [29]	MM22	65.25	83.12	89.00
IITL (T→R)*	-	**67.13**	**84.60**	**90.37**
UNIReID (T→R)*	-	**68.71**	**85.35**	**90.84**

Table 4. Comparison with the state-of-the-arts on Tri-ICFG-PEDES dataset. Rank (R) at k accuracy (%) is reported. $*$ indicates that the original gallery set is used for testing.

Methods	Venue	R1	R5	R10
CMPM/C [46]	ECCV18	43.51	65.44	74.26
SCAN [15]	ECCV18	50.05	69.65	77.21
Dual Path [49]	TOMM20	38.99	59.44	68.41
MIA [22]	TIP20	46.49	67.14	75.18
ViTAA [35]	ECCV20	50.98	68.79	75.78
IVT [30]	Arxiv22	56.04	73.60	80.22
LGUR [29]	MM22	59.02	75.32	81.56
IITL (T→R)*	-	**58.36**	**75.97**	**82.32**
UNIReID (T→R)*	-	**61.28**	**77.40**	**83.16**

Table 5. Comparison with the state-of-the-arts on Tri-RSTPReid dataset. Rank (R) at k accuracy (%) is reported. $*$ indicates that the original gallery set is used for testing.

Methods	Venue	R1	R5	R10
DSSL [50]	MM21	32.43	55.08	63.19
IVT [30]	Arxiv22	46.70	70.00	78.80
LBUL+BERT [37]	MM22	45.55	68.20	77.85
CAIBC [36]	MM22	47.35	69.55	79.00
IITL (T→R)*	-	**57.30**	**78.05**	**86.10**
UNIReID (T→R)*	-	**60.25**	**79.85**	**87.10**

AFL [24], LMDI [12], and SketchTrans [2]. Our cross-domain sketch-based retrieval (*i.e.*, S→R) using a simple single encoder obtains an accuracy better than most supervised training methods that design a complex feature learning framework. With the multi-modal information fusion (*i.e.*, T+S→R), our UNIReID achieves a significant performance improvement, outperforming the latest SketchTrans [2] by 6.8% in Rank-1 accuracy. Moreover, we visualize some retrieval results on the PKU-Sketch dataset in Fig 4. From these results, we conclude that our UNIReID model is well suited for professional hand-drawn sketches, as it is robust to variations in sketch style. It also demonstrates that the fusion of multi-modal data enhances the cross-domain generalization performance.

Table 6. Performance analysis (%) of cross-domain generalization on PKU-Sketch dataset.

Methods	PKU-Sketch				
	R1	R5	R10	mAP	mINP
CD-AFL [24]	34.00	56.30	72.50	-	-
LMDI [12]	49.00	70.40	80.20	-	-
SketchTrans [2]	84.60	94.80	98.20	-	-
UNIReID (T→R)	76.80	93.20	96.20	80.57	77.83
UNIReID (S→R)	69.80	88.60	95.80	72.97	68.25
UNIReID (T+S→R)	**91.40**	**98.80**	**99.80**	**91.76**	**88.97**

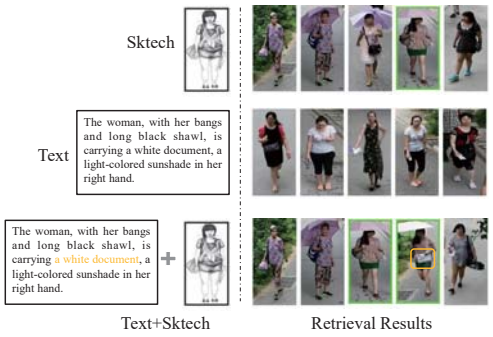

Figure 4. Visualization of qualitative retrieval results of Top-5 for three descriptive queries on the PKU-Sketch dataset.

6. Conclusion

In this paper, we first explore the modality-agnostic person re-identification to deal with descriptive query uncertainty in real scenarios. Also, we construct three multi-modal datasets to assist with the study of this issue. Specifically, we design a unified person re-identification architecture with a simple dual-encoder to handle cross-modality and multi-modality retrieval tasks. In order to learn the modality-shared information for multi-modal data, a task-specific modality learning scheme and task-aware dynamic training strategy are introduced to supervise and balance the training of individual tasks. Moreover, we propose a cross-modality interaction to align the sketch and text representations for modality information fusion. Extensive experimental results on three multi-modal datasets demonstrate the effectiveness of the proposed method.

Acknowledgement. This work is partially supported by the National Natural Science Foundation of China under Grants (62176188, 61771180), the Key Research and Development Program of Hubei Province (2021BAA187, 2022BCA009), the Special Fund of Hubei Luojia Laboratory (220100015), Zhejiang Lab (NO.2022NF0AB01), and the CAAI-Huawei MindSpore Open Fund.

References

[1] Mindspore, https://www.mindspore.cn/, 2020. 6

[2] Cuiqun Chen, Mang Ye, Meibin Qi, and Bo Du. Sketch transformer: Asymmetrical disentanglement learning from dynamic synthesis. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4012–4020, 2022. 1, 2, 3, 4, 6, 7, 8

[3] Cuiqun Chen, Mang Ye, Meibin Qi, Jingjing Wu, Jianguo Jiang, and Chia-Wen Lin. Structure-aware positional transformer for visible-infrared person re-identification. *IEEE Transactions on Image Processing*, 31:2352–2364, 2022. 3

[4] Cuiqun Chen, Mang Ye, Meibin Qi, Jingjing Wu, Yimin Liu, and Jianguo Jiang. Saliency and granularity: Discovering temporal coherence for video-based person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(9):6100–6112, 2022. 1, 2, 7

[5] Feilong Chen, Xiuyi Chen, Shuang Xu, and Bo Xu. Improving cross-modal understanding in visual dialog via contrastive learning. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 7937–7941, 2022. 3

[6] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *Proceedings of the European Conference on Computer Vision*, pages 104–120, 2020. 3, 4

[7] Zefeng Ding, Changxing Ding, Zhiyin Shao, and Dacheng Tao. Semantically self-aligned network for text-to-image part-aware person re-identification. *arXiv preprint arXiv:2107.12666*, 2021. 2, 3, 6

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the International Conference on Learning Representations*, 2020. 4, 6

[9] Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Pengchuan Zhang, Lu Yuan, Nanyun Peng, et al. An empirical study of training end-to-end vision-and-language transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18166–18176, 2022. 3, 4

[10] Chenyang Gao, Guanyu Cai, Xinyang Jiang, Feng Zheng, Jun Zhang, Yifei Gong, Pai Peng, Xiaowei Guo, and Xing Sun. Contextual non-local alignment over full-scale representation for text-based person search. *arXiv preprint arXiv:2101.03036*, 2021. 3

[11] Anjith George and Sebastien Marcel. Cross modal focal loss for rgbd face anti-spoofing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7882–7891, June 2021. 2, 6

[12] Shaojun Gui, Yu Zhu, Xiangxiang Qin, and Xiaofeng Ling. Learning multi-level domain invariant features for sketch re-identification. *Neurocomputing*, 403:294–303, 2020. 3, 8

[13] Ruibing Hou, Hong Chang, Bingpeng Ma, Rui Huang, and Shiguang Shan. Bicnet-tks: Learning efficient spatial-temporal representation for video person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2014–2023, 2021. 1

[14] Ya Jing, Chenyang Si, Junbo Wang, Wei Wang, Liang Wang, and Tieniu Tan. Pose-guided joint global and attentive local matching network for text-based person search. *Association for the Advance of Artificial Intelligence*, 2020. 8

[15] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *Proceedings of the European Conference on Computer Vision*, pages 201–216, 2018. 8

[16] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019. 3, 4

[17] 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. 1, 2, 3, 6

[18] Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. Unimo: Towards unified-modal understanding and generation via cross-modal contrastive learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 2592–2607, 2021. 2

[19] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2):318–327, 2020. 5

[20] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in Neural Information Processing Systems*, 32, 2019. 3

[21] Hao Ni, Jingkuan Song, Xiaopeng Luo, Feng Zheng, Wen Li, and Heng Tao Shen. Meta distribution alignment for generalizable person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2487–2496, 2022. 1

[22] Kai Niu, Yan Huang, Wanli Ouyang, and Liang Wang. Improving description-based person re-identification by multi-granularity image-text alignments. *IEEE Transactions on Image Processing*, 29:5542–5556, 2020. 8

[23] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 5

[24] 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 26th ACM international conference on Multimedia*, pages 609–617, 2018. 1, 3, 7, 8

[25] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*, pages 8748–8763, 2021. 2, 3, 4, 6

[26] Nikolaos Sarafianos, Xiang Xu, and Ioannis A Kakadiaris. Adversarial representation learning for text-to-image matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5814–5824, 2019. 8

[27] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, 2016. 4

[28] Pierre Sermanet, David Eigen, Xiang Zhang, Michael Mathieu, Rob Fergus, and Yann LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. In *Proceedings of the 2nd International Conference on Learning Representations*, 2014. 5

[29] Zhiyin Shao, Xinyu Zhang, Meng Fang, Zhifeng Lin, Jian Wang, and Changxing Ding. Learning granularity-unified representations for text-to-image person re-identification. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 5566–5574, 2022. 1, 3, 7, 8

[30] Xiujun Shu, Wei Wen, Haoqian Wu, Keyu Chen, Yiran Song, Ruizhi Qiao, Bo Ren, and Xiao Wang. See finer, see more: Implicit modality alignment for text-based person retrieval. *arXiv preprint arXiv:2208.08608*, 2022. 8

[31] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019. 3

[32] Jonas Uhrig, Marius Cordts, Uwe Franke, and Thomas Brox. Pixel-level encoding and depth layering for instance-level semantic labeling. In *Proceedings of the German Conference on Pattern Recognition*, pages 14–25, 2016. 5

[33] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017. 4, 6

[34] Chengji Wang, Zhiming Luo, Yaojin Lin, and Shaozi Li. Text-based person search via multi-granularity embedding learning. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pages 1068–1074, 2021. 8

[35] Zhe Wang, Zhiyuan Fang, Jun Wang, and Yezhou Yang. Vitaa: Visual-textual attributes alignment in person search by natural language. In *Proceedings of the European Conference on Computer Vision*, pages 402–420, 2020. 8

[36] Zijie Wang, Aichun Zhu, Jingyi Xue, Xili Wan, Chao Liu, Tian Wang, and Yifeng Li. Caibc: Capturing all-round information beyond color for text-based person retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 5314–5322, 2022. 8

[37] Zijie Wang, Aichun Zhu, Jingyi Xue, Xili Wan, Chao Liu, Tian Wang, and Yifeng Li. Look before you leap: Improving text-based person retrieval by learning a consistent cross-modal common manifold. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 1984–1992, 2022. 8

[38] Bin Yang, Mang Ye, Jun Chen, and Zesen Wu. Augmented dual-contrastive aggregation learning for unsupervised visible-infrared person re-identification. In *Proceedings of the 30th ACM International Conference on Multimedia*, page 2843–2851, 2022. 3

[39] 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*, 2020. 3

[40] Mang Ye, Cuiqun Chen, Jianbing Shen, and Ling Shao. Dynamic tri-level relation mining with attentive graph for visible infrared re-identification. *IEEE Transactions on Information Forensics and Security*, 17:386–398, 2021. 3

[41] Mang Ye, Weijian Ruan, Bo Du, and Mike Zheng Shou. Channel augmented joint learning for visible-infrared recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13567–13576, 2021. 1, 2, 3, 4, 6

[42] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6):2872–2893, 2021. 1, 2, 6

[43] Yajing Zhai, Yawen Zeng, Da Cao, and Shaofei Lu. Trireid: Towards multi-modal person re-identification via descriptive fusion model. In *Proceedings of the 2022 International Conference on Multimedia Retrieval*, pages 63–71, 2022. 1, 2, 3, 4, 6, 7

[44] Shengyu Zhang, Tan Jiang, Tan Wang, Kun Kuang, Zhou Zhao, Jianke Zhu, Jin Yu, Hongxia Yang, and Fei Wu. Devlbert: Learning deconfounded visio-linguistic representations. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 4373–4382, 2020. 3

[45] Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. Contrastive learning of medical visual representations from paired images and text. *arXiv preprint arXiv:2010.00747*, 2020. 4

[46] Ying Zhang and Huchuan Lu. Deep cross-modal projection learning for image-text matching. In *Proceedings of the European conference on computer vision*, pages 686–701, 2018. 8

[47] Zhong Zhang, Haijia Zhang, and Shuang Liu. Person re-identification using heterogeneous local graph attention networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12136–12145, 2021. 1

[48] Kecheng Zheng, Wu Liu, Lingxiao He, Tao Mei, Jiebo Luo, and Zheng-Jun Zha. Group-aware label transfer for domain adaptive person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5310–5319, 2021. 1

[49] Zhedong Zheng, Liang Zheng, Michael Garrett, Yi Yang, Mingliang Xu, and Yi-Dong Shen. Dual-path convolutional image-text embeddings with instance loss. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 16(2):1–23, 2020. 8

[50] Aichun Zhu, Zijie Wang, Yifeng Li, Xili Wan, Jing Jin, Tian Wang, Fangqiang Hu, and Gang Hua. Dssl: Deep surroundings-person separation learning for text-based person retrieval. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 209–217, 2021. 2, 3, 6, 8