

## Highlights

### **From Attributes to Natural Language: A Survey and Foresight on Text-based Person Re-identification**

Fanzhi Jiang, Su Yang, Mark W. Jones, Liumei Zhang

- The most comprehensive survey of text-based person re-identification to date.
- A novel taxonomy for text-based person re-identification from different dimensions.
- Challenges in text-based person Re-ID are analyzed, with future directions given.
- A open-world baseline for text-based pedestrian image generation guided Re-ID.

# From Attributes to Natural Language: A Survey and Foresight on Text-based Person Re-identification<sup>★</sup>

Fanzhi Jiang<sup>a,b</sup>, Su Yang<sup>a,b,\*</sup>, Mark W. Jones<sup>a</sup> and Liumei Zhang<sup>c,d,\*</sup>

<sup>a</sup>*School of Mathematics and Computer Science, Swansea University, Fabian Way, Swansea, SA1 8EN, Wales, UK*

<sup>b</sup>*Computer Vision & Machine Learning Lab, Swansea University, Fabian Way, Swansea, SA1 8EN, Wales, UK*

<sup>c</sup>*School of Computer Science, Xi'an Shiyou University, Dianzi 2nd Road, Xi'an, 710065, Shaanxi, China*

<sup>d</sup>*ChengYin Lab, Xi'an Shiyou University, Dianzi 2nd Road, Xi'an, 710065, Shaanxi, China*

## ARTICLE INFO

### Keywords:

Person Re-identification

Text

Natural Language

Attributes

Diffusion Model.

## ABSTRACT

Text-based person re-identification (Re-ID) is a challenging topic in the field of complex multimodal analysis, its ultimate aim is to recognize specific pedestrians by scrutinizing attributes/natural language descriptions. Despite the wide range of applicable areas such as security surveillance, video retrieval, person tracking, and social media analytics, there is a notable absence of comprehensive reviews dedicated to summarizing the text-based person Re-ID from a technical perspective. To address this gap, we propose to introduce a taxonomy spanning *Evaluation*, *Strategy*, *Architecture*, and *Optimization* dimensions, providing a comprehensive survey of the text-based person Re-ID task. We start by laying the groundwork for text-based person Re-ID, elucidating fundamental concepts related to attribute/natural language-based identification. Then a thorough examination of existing benchmark datasets and metrics is presented. Subsequently, we further delve into prevalent feature extraction strategies employed in text-based person Re-ID research, followed by a concise summary of common network architectures within the domain. Prevalent loss functions utilized for model optimization and modality alignment in text-based person Re-ID are also scrutinized. To conclude, we offer a concise summary of our findings, pinpointing challenges in text-based person Re-ID. In response to these challenges, we outline potential avenues for future open-set text-based person Re-ID and present a baseline architecture for text-based pedestrian image generation guided re-identification (*TBPGR*).

## 1. Introduction

The demands in public safety and the subsequently installed surveillance networks make the manual tracking and identifying individuals increasingly challenging. Automatic person re-identification (Re-ID) has emerged as a popular research area in computer vision to address this issue. Also known as person search [1], person Re-ID specifically refers to recognizing and tracking specific individuals using non-overlapping images and videos captured by cameras, determining whether a specific query person appears persistently or momentarily across different times and locations [2]. It finds important applications in scenarios like searching for characters in movies, locating missing children, and tracking criminals [3]. The source data collection for these applications is usually from CCTV footage on the street. Given the constraints of the equipment and the environment, a number of challenges arise. These challenges include lighting variations, occlusions, background changes, pose variations, and differences in camera resolutions. To address the limitations of the source data, there are also

some works that propose to construct large-scale diverse synthetic datasets to enhance the performance of the re-identification task, including data collecting and labeling [4] [5], person generation [6]. A typical person Re-ID system involves conducting image queries and searching for corresponding individuals in a gallery of images or a video pool. It involves extracting features directly from pedestrian appearance images and performing direct matching, then ranking against existing appearance images in the gallery. These scenarios assume that visual examples of individual identities are always available as queries. However, some special cases where the visualization samples of personal identity are not available, we perform retrieval based on textual descriptions only, which is called text-based person Re-ID, as shown in Figure 1.

### 1.1. Attribute-based Person Re-ID

Text-based person Re-ID is a special form of person Re-ID that, instead of using image data, relies on descriptions to reflect a person's appearance [7]. In the task of text-based person Re-ID, a natural language description and an image are provided as input, and the output is the identification of the person matching the description [8]. There has been a desire to perform high-level semantic person search using free-form natural language descriptions. However, early research initially started with the attribute-based person Re-ID [9, 10, 11, 12]. Attribute-based methods (Pedestrian Attributes Recognition, PAR) stem from the high correlation between attributes and pedestrian images. Pedestrian attributes such as gender, age, clothing type, clothing color,

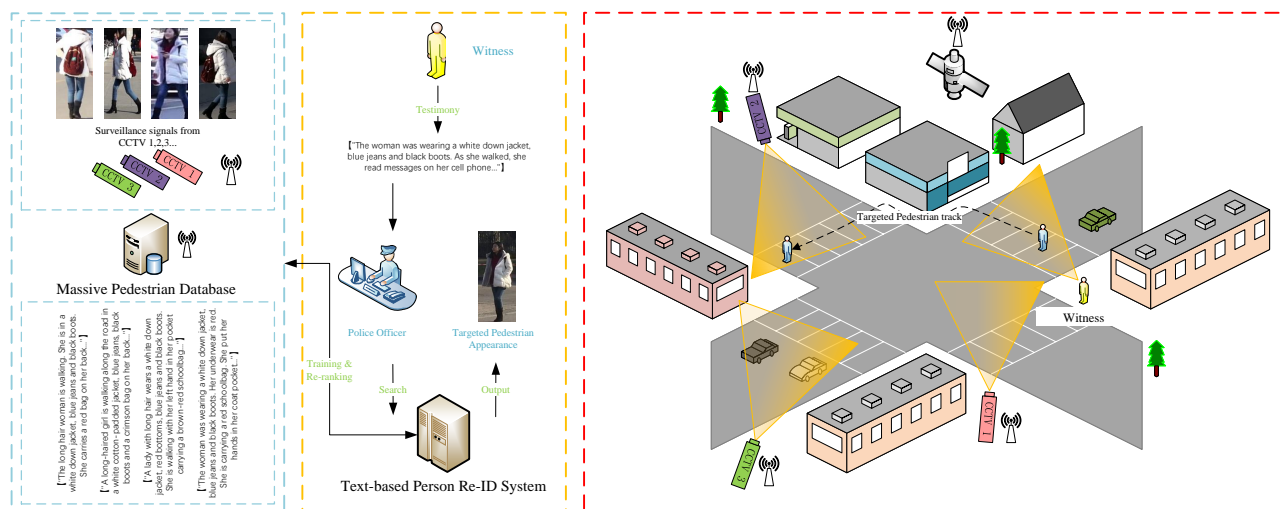
<sup>★</sup>This document is the results of the research project funded by The Engineering and Physical Sciences Research Council of UK Research and Innovation (UKRI).

All codes and datasets are available at <https://github.com/fonsjiang/TBPGR>.

\*Co-corresponding authors

✉ f. jiang.2250053@swansea.ac.uk (F. Jiang); su.yang@swansea.ac.uk (S. Yang); m.w.jones@swansea.ac.uk (M.W. Jones); zhangliumei@xsyu.edu.cn (L. Zhang)

ORCID(s): 0000-0001-7229-9732 (F. Jiang); 0000-0002-6618-7483 (S. Yang); 0000-0001-8991-1190 (M.W. Jones); 0000-0002-1834-5424 (L. Zhang)



**Figure 1:** Conceptual diagram of text-based person Re-ID. Given a textual description of a target person collected from a street witness, a monitor uses the model aimed at retrieving the corresponding person image from a given database of images collected from street CCTV.

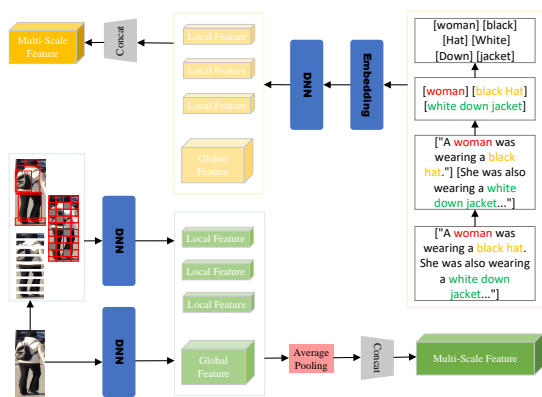
etc. are arranged and combined to make the pedestrian identity appear distinguishable [13]. Extracting attribute information from pedestrian images can be achieved by using pre-trained classifiers or attribute detectors. These classifiers or detectors can recognize and locate different attributes (body part) in the images. These extracted pedestrian attribute information needs to be encoded into a feature vectors pattern matching. This often involves transforming discrete attributes into continuous feature representations, common methods include one-hot encoding and embedding encoding [12]. During the recognition process, matching is conducted by comparing the attribute feature vectors of query images and gallery images. Common matching methods include computing distances or similarities between feature vectors (such as Euclidean distance, cosine similarity, etc.). Finally, based on the results of attribute matching, the final recognition result is determined through fusion of matching scores from different attributes or by adopting decision strategies. The attribute-based person Re-ID approach improves recognition accuracy and retrieval interpretability. This is because these attributes can be considered as high-level semantic information that is robust to viewpoint changes and different viewing conditions [14].

Previous work on attribute-based person retrieval has attempted to reduce modality gaps by aligning each person category and its corresponding images in a joint embedding space through modal adversarial training [15, 16], or by enhancing the expressive power of embedding vectors for person categories and images hierarchically [12]. While these pioneering studies reveal important but less explored methods for person retrieval, there remains significant room for further improvement. Firstly, due to their adversarial learning strategy, they exhibit instability and high computational cost during training [2, 17]. For other, they incur high inference costs due to the need for additional networks to match high-dimensional embedding vectors [12]. More

importantly, these methods treat person categories as independent labels of person images and overlook their relationships, such as how many attributes differ between them, despite the potential rich supervisory signals for learning better representations of person categories and images. Dong et al. [18] address this issue by capturing rich information of two modalities through hierarchical embedding. However, their model entails high computational complexity as it computes high-dimensional embeddings and deploys additional networks for matching. Ref. [16] and [2] learn a joint embedding space where person categories and images are directly matched. To bridge the modality gap, their embedding space is trained in a modal adversarial manner; however, due to the nature of min-max optimization, this often leads to unstable and slow convergence. Furthermore, all these methods suffer from a limitation where person categories are treated as separate labels and the important relationships among them are overlooked. In the joint embedding space of the two modalities, the loss pulls images of the same person category closer to achieve modal alignment. Jeong et al. [19] introduces a novel loss, Adaptive Semantic Margin Regularizer (ASMR), for learning cross-modal embeddings in the context of attribute-based person retrieval.

## 1.2. Natural Language-based Person Re-ID

Attribute-based person Re-ID [20, 21] provides a method for person re-identification using predefined attributes. However, this approach has limitations in certain scenarios. For example, the PETA dataset [12] defines 61 binary attributes, 4 multi-class attributes, and hundreds of phrases to describe appearance. Rearranging or flipping these attributes during retrieval can significantly degrade performance. Additionally, annotating large-scale person image datasets with such attributes is time-consuming and costly, even when the attribute set is limited. This indicates certain weaknesses in robustness.



**Figure 2:** Multi-scale feature fusion: The figure shows the basic operations regarding the fusion of image and text multi-scale features in person re-identification.

Unlike these limitations, natural language descriptions offer an unstructured alternative that is more flexible and robust. While most attributes are noun phrases, natural language can include higher-level semantic expressions and complex events. In principle, natural language (NL) processing techniques can describe any attribute with no restrictions on word choice or length, capturing unique details. Furthermore, natural language descriptions can include indicators of certainty or ambiguity, preserving as much information as possible and supporting multi-class attribute assignments. In this study, we refer to both attribute and NLP-based person re-identification as text-based person Re-ID.

Li et al. [7] introduced the first natural language-based person Re-ID dataset, which evaluated early cross-modal models and proposed a recurrent neural network with gated neural attention. Later, Zheng et al [22] conducted the first attempt at language-based pedestrian search using unsegmented whole images, advancing the field in more complex and realistic scenarios. Subsequent studies mainly focused on fine-grained visual-text matching. Ref. [23] emphasized the importance of mining fine-grained local features to learn distinctive local representations. However, these methods often overlooked the associations between image regions and words with different semantic granularities [24, 21], leading to ambiguous embeddings. Discriminative parts of a pedestrian often appear at different granularities, and ignoring this can degrade matching performance. To address these issues, incorporating words as guidance to align with relevant image regions has become an effective strategy. Niu et al. [25] proposed the Multi-level Image-Text Alignment (MIA) model, which introduced cross-modal adaptive attention at three different granularities. However, integrating irrelevant words may mislead the model and exacerbate embedding ambiguity.

Cross-domain embedding techniques have been a key research focus in the Re-ID community. Li et al. [20] employed a shared attention mechanism to learn cross-modal embeddings, improving matching performance. Ref. [26]

enriched annotations to extract meaningful image regions independently, though such regions may not always be available in real-world scenarios. Zhang et al. [27] proposed a cross-modal projection loss, widely used for learning discriminative image and text embeddings. Additionally, Chen et al. [28] introduced a block-word matching model with an adaptive threshold mechanism to establish both global and local image-language associations for better semantic consistency. Sarafianos et al. [29] and Liu et al. [30] proposed adversarial learning-based graph models to enhance the robustness of cross-domain representations. To address variations in style and environment, Zhu et al. [31] introduced an environment-person separation method under exclusive constraints and proposed the Real-Scene Text-based Person Re-Identification (RSTPReID) dataset for Re-ID. Wu et al. [32] designed a color-driven dual-task framework, including image colorization and text completion, establishing fine-grained cross-modal associations through color-based reasoning for the first time. Yan et al. [33] proposed a dictionary learning algorithm based on matrix decomposition to mitigate the effects of style and pedestrian pose on cross-domain Re-ID. Ref. [34] developed the Cross-modal Co-occurrence Attributes Alignment (C2A2) method to handle noise, reducing interference from irrelevant pedestrian elements.

However, traditional text-based person Re-ID methods heavily rely on identity annotations, making the manual labeling process both time-consuming and expensive. To address the issue of limited data learning, Han et al. [35] proposed transferring knowledge from existing large-scale coarse-grained datasets to compensate for the lack of training data. Zhao et al. [36] considered a more practical setup by introducing weak supervision into text-based person Re-ID, where only text-image pairs are available without identity annotations during training. Yang et al. [37] leveraged pre-trained diffusion models to create a large multi-attribute and language search dataset (MALS), exploring the feasibility of pretraining on both attribute recognition and text-image matching tasks. Many more works [38, 39, 40, 41, 42], all shifted to new feature extraction and knowledge transfer paradigms with the rise of the large-scale multimodal model CLIP, marking a new stage of development in text-based person Re-ID [43, 44, 45].

In summary, significant progress has been made in text-based person Re-ID through techniques such as multi-scale learning, cross-domain embedding, noise reduction, and large model pretraining. Future research may focus on domain adaptation, open set retrieval, and privacy preservation to better address open-world challenges.

### 1.3. Person Re-ID previous survey

In order to provide researchers with a comprehensive understanding of the current status and research directions in the domain of person Re-ID, we conducted an in-depth investigation of text-based person Re-ID methodologies and synthesized recent research achievements. Prior to this effort, several researchers have conducted reviews in the field

of person Re-ID [3, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63] and we summarize the primary contributions of these studies in the Table 3. These surveys can be broadly categorized into the following groups:

**Surveys on Traditional Methods:** Early surveys introduced the foundational challenges in person Re-ID and traditional approaches to address them [46, 48].

**Surveys on Deep Learning Techniques:** With the advent of deep learning, several surveys focused on summarizing deep learning methods for pedestrian Re-ID [49, 50, 53, 58, 59].

**Metric Learning Approaches:** Other surveys delved into metric learning-based methods for person Re-ID [64, 54, 1].

**Representation Learning Taxonomies:** Studies such as [52, 1, 58, 59, 64] presented taxonomies based on representation learning for person Re-ID.

**Image/Video-based person Re-ID:** Some studies investigated the available image/video-based person Re-ID datasets, evaluation metrics and methods used in the field [57, 62, 58, 59, 61].

**Intra-Modality person Re-ID:** Surveys like [63] provided overviews of visible-infrared cross-modal Re-ID studies. [3] reviewed cross-modality based person Re-ID research from four aspects: sketch, text, low resolution, and infrared.

However, existing surveys on person Re-ID do not provide systematic categorization, detailed analysis, or comprehensive discussion of text-based methods, as illustrated in Figure 3. To address these gaps, this paper conducts an in-depth analysis of recent advancements in text-based person Re-ID, covering key dimensions such as datasets, strategies, architectures, and optimization techniques. Besides, The survey also highlights the strengths and limitations of existing methods, providing insights for future research. Furthermore, we propose a novel baseline architecture for text-based pedestrian image generation guided re-identification (TBPGR), designed to address challenges in open-world scenarios. This paradigm bridges the gap between text and image modalities, enabling more robust and open Re-ID in practical applications.

To provide a clear roadmap for this survey, the structure of this paper is summarized as follows. Section 2 reviews publicly available datasets and evaluation metrics commonly used in text-based person Re-ID. Section 3 discusses various strategies for feature extraction and cross-modal alignment. Section 4 examines the architectures of deep learning models applied in this domain. Section 5 focuses on optimization methods, including commonly used loss functions. Section 6 analyzes the challenges faced by current methods. Finally, Section 7 outlines potential future directions and introduces a novel baseline architecture TBPGR for text-based pedestrian image generation guided re-identification.

## 2. Evaluation

The evaluation of text-based person Re-ID tasks, typically involves publicly available benchmark datasets and common performance metrics. This provides a standardized testing platform, enabling a quantified and objective comparative assessment of the effectiveness of different experimental algorithms.

### 2.1. Datasets

#### 2.1.1. Attribute-based Person Re-ID Datasets

Attribute-based Person Re-ID often involves the utilization of semantic attributes that describe appearance characteristics of individuals, such as gender, age, clothing type, clothing texture, and clothing color [71]. However, these attributes frequently encompass low-level semantic attributes that can be directly associated with image regions, such as clothing type, clothing texture, and clothing color, as well as high-level semantic attributes that directly correspond to people, such as gender and age. By leveraging these attributes, the negative impact of variations in viewpoint and appearance on person Re-ID can be mitigated, resulting in more reliable recognition and tracking performance [19]. To facilitate research in Attribute-based Person Re-ID, researchers have initiated the construction of several datasets specifically tailored for this purpose. Notable publicly available datasets in this domain include PETA [12], UAV-Human [69], RAP [66], RAP2.0 [67], PA-100K [68], Market1501-Attribute [11], and DukeMTMC-Attribute [11].

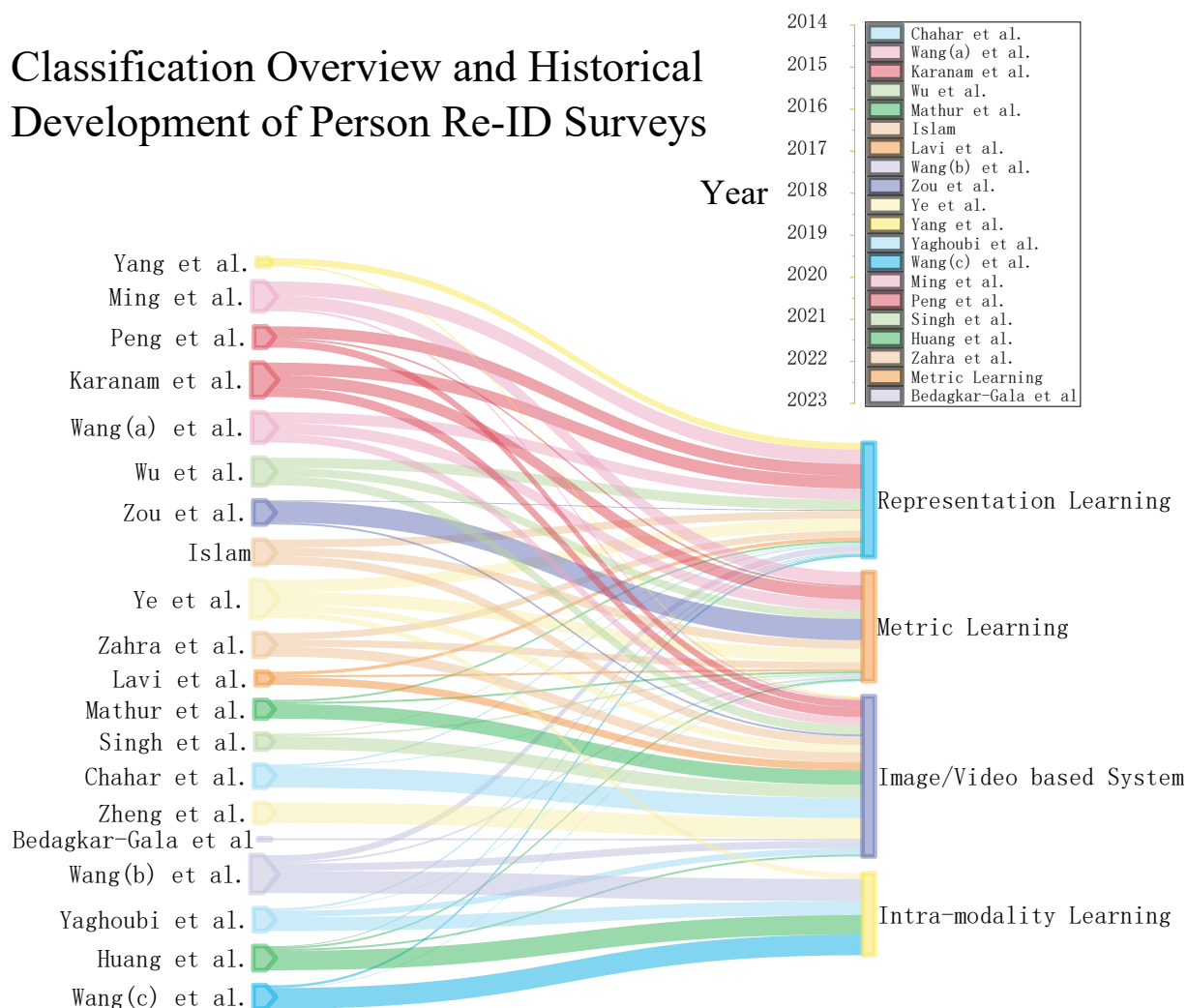
These datasets typically consist of a substantial collection of images capturing individuals from surveillance cameras, each accompanied by attribute labels. The dataset creation process generally involves selecting a subset of images from existing Person Re-ID datasets. Subsequently, attribute labels are assigned to each image either manually or through automated annotation. Ultimately, the annotated images along with their corresponding attribute labels are organized into a unified dataset, which serves as a resource for researchers in the field, as shown in Table 1.

#### 2.1.2. Natural Language-based Person Re-ID Datasets

Publicly available NL-based person Re-ID datasets have significant practical implications for benchmarking, algorithm improvement, security and surveillance applications, privacy protection, and advancing multimodal research. Currently, there are three large publicly available datasets, CUHK-PEDES, RSTPREid, and ICFG-PEDES. The details of these datasets will be described.

**CUHK-PEDES** CUHK-PEDES is the first description dataset introduced by The Chinese University of Hong Kong [7], designed for training and evaluating person description and image retrieval tasks in computer vision. The dataset was collected by aggregating person images from five existing and well-known person Re-ID evaluation datasets, namely CUHK03 [72], Market-1501 [47], SSM [73], VIPER [74], CUHK01 [75], and subsequently annotated with descriptive

# Classification Overview and Historical Development of Person Re-ID Surveys



**Figure 3:** The figure shows the development of the main categorical dimensions and surveys of Person Re-ID throughout history. Overall, the large figure on the left represents the relevance of influential Person Re-ID surveys and their categorical dimensions, with the line font implicitly indicating their phase weights. The smaller figure on the top right is a more visual representation of the temporal relationships between the emergence of these Person Re-ID surveys.

**Table 1**

Summary of public datasets for attribute-based and natural language-based person Re-ID. The datasets are summarized in terms of dimensions such as dataset, number of pedestrians, attributes/description, and data source, respectively.

|                 | Dataset               | Pedestrians        | Attribute/Description    |             | Source         |
|-----------------|-----------------------|--------------------|--------------------------|-------------|----------------|
|                 |                       |                    | Binary                   | Multi-class |                |
| Attribute-based | PETA [65]             | 19,000             | 61                       | 4           | Indoor         |
|                 | RAP [66]              | 41,585             | 69                       | 3           | Indoor         |
|                 | RAP V2.0 [67]         | 84,928             | 69                       | 3           | Outdoor        |
|                 | PA-100K [68]          | 100,000            | 26                       | –           | Outdoor        |
|                 | Market1501-Attr. [11] | 32,668             | 26                       | 1           | Outdoor        |
|                 | DukeMTMC-Attr. [11]   | 34,183             | 23                       | –           | Outdoor        |
|                 | UAV-Human [69]        | 22,263             | 7                        | –           | Outdoor        |
| NL-based        | <b>Dataset</b>        | <b>Pedestrians</b> | <b>Sentences × Words</b> |             | <b>Source</b>  |
|                 | CUHK-PEDES [7]        | 34,054             | 2                        | >23         | Outdoor+Indoor |
|                 | RSTPREid [31]         | 20,505             | 2                        | >25.8       | Outdoor+Indoor |
|                 | ICFG-PEDES [70]       | 54,522             | 3                        | >37.2       | Outdoor+Indoor |

text by crowdsourced workers from Amazon Mechanical Turk (AMT). It consists of over 40,206 images with 80,412 person descriptions, each containing at least 23 words, across 13,003 unique person identities. The training set includes 34,054 images, 11,003 unique persons, and 68,108 textual descriptions. The validation set contains 3,078 images, 1,000 unique persons, and 6,158 textual descriptions, while the test set comprises 3,074 images, 1,000 unique persons, and 6,156 textual descriptions. The individuals in the dataset were captured in street and public places in the Hong Kong area. CUHK-PEDES provides diverse person descriptions encompassing appearance, actions, postures, and interactions, with each person description linked to the corresponding image ID. The release of the CUHK-PEDES dataset promotes interdisciplinary research between computer vision and natural language processing, providing researchers in the fields of person description and image retrieval with the first benchmark dataset.

**RSTPReid** The RSTPReid dataset is a person Re-ID dataset introduced by Nanjing University of Science and Technology [31], designed for training and evaluating person Re-ID tasks in computer vision. To address the issue in the CUHK-PEDES dataset where each specific pedestrian is captured by the same camera under the same time-space conditions, which does not reflect real-world scenarios, the authors constructed the Real Scenarios Text-based Person Reidentification (RSTPReid) dataset based on MSMT17 [76]. The RSTPReid dataset contains 20,505 images of 4,101 individuals captured by 15 independent cameras from different viewpoints, lighting conditions, locations, and weather conditions. Each individual has 5 corresponding images taken by different cameras, and each image is accompanied by 2 text descriptions with no fewer than 23 words. The training set includes 3,701 identities, the validation set has 200 identities, and the test set also consists of 200 identities. The dataset includes a relatively small number of pedestrian identities but covers both indoor and outdoor scenarios. Each pedestrian appears in multiple images with different shooting angles and lighting conditions, making this dataset more challenging and realistic.

**ICFG-PEDES** The Identity-Centric and Fine-Grained Person Description Dataset (ICFG-PEDES) is a person description dataset introduced by South China University of Technology [70], which also serves as a new benchmark dataset for research in person description and image retrieval domains. Similar to CUHK-PEDES, this dataset contains a large number of person descriptions paired with corresponding image IDs. ICFG-PEDES consists of 54,522 pedestrian images of 4,102 distinct identities, all collected from the MSMT17 database [76]. The training set comprises 34,674 image-text pairs of 3,102 pedestrians, while the test set contains 19,848 image-text pairs of 1,000 pedestrians. Each image is associated with only one text description, with an average of 37.2 words per description. Compared to the CUHK-PEDES dataset, ICFG-PEDES emphasizes

more on fine-grained person descriptions and reduces some irrelevant action and background information. Additionally, it addresses the issue of consistent backgrounds in the former dataset by emphasizing more on appearance variations. Consequently, the ICFG-PEDES dataset can be employed for more challenging image retrieval and person description tasks.

## 2.2. Metrics

### 2.2.1. Mean Average Precision(mAP)

For each query  $i$ , we define a precision  $P_i(j)$  as the proportion of correct matches in the top  $j$  matches, and then for each positive instance, we calculate the average of the proportions of all positive instances before this one, which is the Average Precision  $AP_i$ :

$$AP_i = \frac{1}{M_i} \sum_{j=1}^{M_i} P_i(j) \cdot I(\text{rank}_i = j). \quad (1)$$

Where  $M_i$  is the number of positive instances for query  $i$ , and  $\sum$  is the summation over all positive instances. The symbol  $I(\text{rank}_i = j)$  usually denotes an indicator function. The indicator function takes the value of 1 under certain conditions, otherwise it takes the value of 0. Specifically:

$$I(\text{rank}_i = j)$$

indicates that the indicator function  $I(\text{rank}_i = j)$  takes the value of 1 when the first  $i$  query's first  $j$  result is a correct match, otherwise it takes the value of 0. In other words, it determines whether or not there is a correct match in the ranked position of  $j$ . Then, mAP is the average of the Average Precision of all queries:

$$mAP = \frac{1}{N} \sum_{i=1}^N AP_i. \quad (2)$$

Where  $N$  is the total number of queries, and  $\sum$  is the summation over all queries.

### 2.2.2. Rank-N Accuracy

For each query  $i$ , we define an indicator function  $I(\text{rank}_i \leq n)$ , which is 1 if the correct match is within the top  $n$  matches, and 0 otherwise. Then, Rank- $n$  accuracy is the average of this function:

$$\text{Rank} - n = \frac{1}{N} \sum_{i=1}^N I(\text{rank}_i \leq n). \quad (3)$$

Where  $N$  is the total number of queries, and  $\sum$  is the summation over all queries.

## 3. Strategy

In the task of text-based person Re-ID, image data is typically captured through CCTV in open environments [1].

This data not only contains key descriptive features of individuals but also includes a significant amount of background noise. Early works [7, 27, 33, 28] typically directly extract global image text features for simple cross-modal matching. However, due to the uniqueness of the person Re-ID task, which differs from existing image-text retrieval [22, 17], the categories to be retrieved all belong to highly similar individuals. Identification based solely on distinguishing features such as clothing and posture increases the challenges of the text-based person Re-ID task.

### 3.1. Stripe Segmentation

Stripe Segmentation is a widely used technique in person Re-ID to mitigate the impact of pose variations, occlusions, and lighting conditions on performance [25, 70]. The core idea is to partition pedestrian images into multiple vertical stripes, each representing a subregion of the image, such as the head, upper body, lower body, or feet. This method effectively captures local features by dividing images into local regions, enhancing recognition accuracy [77]. Local features like texture, color, and shape can be extracted from each stripe, facilitating better differentiation between different parts of pedestrians. Stripes of varying heights correspond to different scales of pedestrian parts, providing richer scale information to address scale variation issues in images.

Several works employ Stripe Segmentation to achieve multimodal local alignment supervision. For instance, Ref. [25, 78] suggest that body parts are evenly arranged in images, using Stripe Segmentation as a guiding model. Ding et al. [70] designed a multi-view non-local network to capture relationships between body parts, establishing better correspondences between body parts and noun phrases. Li et al. [77] vertically divided character images into multiple regions using overlapping slices and key-point-based slices. Similarly, Ref. [25] emphasized local-local alignment by employing Stripe Segmentation in their Bidirectional Fine-Grained Matching (BFM) module to match visual body parts with noun phrases. However, this strategy can be fragile and sensitive to conditional changes. For example, image samples may not always contain a complete human body, and the head may not consistently appear in the first stripe, which significantly impacts the robustness of these methods.

### 3.2. Multi-scale Fusion

In pedestrian retrieval tasks, the small inter-class variance in images and descriptions necessitates comprehensive information to coordinate visual and textual clues across all scales. Multi-scale feature fusion techniques address this need by combining features at different levels [79], as illustrated in Figure 2.

#### 3.2.1. Attribute Segmentation

Some methods focus on decomposing visual features into attribute-specific subspaces to enhance alignment with textual descriptions. Wang et al. [24] proposed the Visual Text Attribute Alignment model (ViTAA), which uses an attribute segmentation layer to divide pedestrian images into parts such as full body, head, upper clothes, lower clothes,

shoes, and bag. Corresponding textual phrases are extracted for each segmented part, facilitating fine-grained alignment between images and text. Similarly, Liu et al. [30] introduced the Image Structure Graph Network (A-GANet), which employs residual modules for visual feature extraction and graph attention convolutional layers for high-level semantic representation. It captures relationships between different pedestrian parts using graph convolutional networks, enhancing the structured representation of visual features.

#### 3.2.2. Multi-Granularity Alignment

Several approaches focus on aligning image and text features at multiple granularities. Niu et al. [25] introduced the Multi-Granularity Image-Text Alignment (MIA) framework and used three modules to solve the cross-modal fine-grained alignment problem. Zheng et al. [21] introduced a Hierarchical Adaptive Matching model that captures fine-grained image-text correspondences at the word, phrase, and sentence levels, facilitating precise feature alignment. Wang et al. [80] developed the Multi-Granularity Embedding Learning (MGEL) model, which enhances retrieval performance by extracting embeddings from human body images at various spatial scales. Ji et al. [78] proposed the Asymmetric Cross-Scale Alignment (ACSA) method, which combines global text representations and local phrase representations while segmenting visual features into key regions (e.g., head, torso, limbs). This partitioning strategy preserves critical details for fine-grained matching with minimal computational cost.

#### 3.2.3. Implicit Aggregation Alignment

Other methods focus on aligning features without explicit supervision or complex interactions. Yan et al. [81] proposed an Implicit Local Alignment module that adaptively aggregates image and text features into modality-shared semantic topic centers. This approach implicitly learns fine-grained correspondences between images and text without additional supervision, supplemented by global alignment. Gao et al. [82] introduced the Non-Local Alignment on Full Scale (NAFS) method, which adaptively aligns image and text features at all scales. It uses a ladder network structure to extract full-size image features with enhanced locality and employs a language model with local constraint attention to obtain description representations at different scales. A context non-local attention mechanism is applied to discover potential alignments simultaneously across scales.

#### 3.2.4. Local-Global Correlation

Some approaches tackle the problem of local-global correlation in embedding spaces. Wang et al. [83] designed the Divide and Merge Embedding (DME) learning framework for text-based person search. It models relationships between local parts and global embeddings, merging local details into the global representation to improve alignment. Ref. [84] proposed constructing features for person-text-image matching by placing features with the same semantics in the same spatial positions. This strategy achieves semantic

consistency and interpretability of global features, addressing misalignment issues caused by aggregating aligned local features into global ones.

### 3.2.5. Multi-Branch Representation

Multi-branch representations have been used to enable text-adaptive matching of visual local representations. Chen et al. [8] adopted a multi-branch representation in the learning path and proposed a multi-stage cross-modal matching strategy. This approach eliminates modality gaps from low-level, local, and global features, gradually narrowing the feature gap between image and text domains. Li et al. presented the Joint Label and Feature Alignment Framework (TFAF) [85] to reduce inter-modal and intra-class gaps. It constructs a dual-path feature learning network for feature extraction and alignment, uses a text generation module to generate label sequences from visual features for label alignment, and introduces a fusion interaction module with a multi-stage feature fusion strategy to eliminate modality heterogeneity.

### 3.2.6. Summary

In summary, multi-scale fusion techniques enhance pedestrian retrieval by coordinating visual and textual clues across different scales and granularities. While methods that apply target detection or additional branch networks to detect significant regions and extract local features offer accuracy advantages, they often incur higher computational costs due to increased network complexity.

## 3.3. Attention Mechanism

Cross-modal alignment is challenging for fine-grained matching between text and images. Attention mechanisms enhance feature extraction by capturing discriminative features related to language descriptions and visual appearances. They intuitively aid in understanding and controlling the alignment process without relying on external cues. Attention can be divided into hard attention [86], [21] and soft attention according to broad strategies. We further meticulously categorize the soft attention used in the literature into spatial [80], [87] and channel attention [81], [88], mixed attention [89], [83], non-local and contextual attention [90], [82] and cross-modal attention [91], [40], [92].

### 3.3.1. Spatial and Channel Attention

Spatial and channel attention methods focus on capturing local discriminative features and maximizing complementary information from different scales. Wang et al. [80] proposed an attention-based deep neural network that captures multiple attention features from low-level to semantic layers for fine-grained pedestrian representation. Liu et al. [87] utilized a multi-head attention module to extract embeddings of different granularities from the text stream, employing adaptive filtering to obtain fine-grained features.

Yan et al. [81] described an image-specific information suppression module that utilizes relationally guided localization and channel attention filtering to suppress background and ambient factors and achieve information alignment between text and images. Ref. [88] utilized channel attention to

focus on the person in an augmented input that connected the original 3 image channels with an additional 14 pose confidence maps to augment the visual representation. However, Spatial and channel attention are typically modeled within a specific feature space and may have difficulty handling complex multimodal inputs or semantic relationships.

### 3.3.2. Mixed Attention

Mixed attention methods combine different attention types to enhance feature extraction. Li et al. [89] proposed a cubic attention convolutional neural network that combines spatial and channel attention to maximize complementary information from different scales, addressing cross-modal alignment challenges. Wang et al. designed a Feature Division Network (FDN) that embeds inputs into  $K$  locally guided semantic representations using self-attention [83], representing different person parts, and then merges them into a compact global embedding. Yang et al. employed an architecture containing text images and attribute branches to achieve stronger feature fusion through mixed attention is combining complementary information between features [37].

### 3.3.3. Non-Local and Contextual Attention

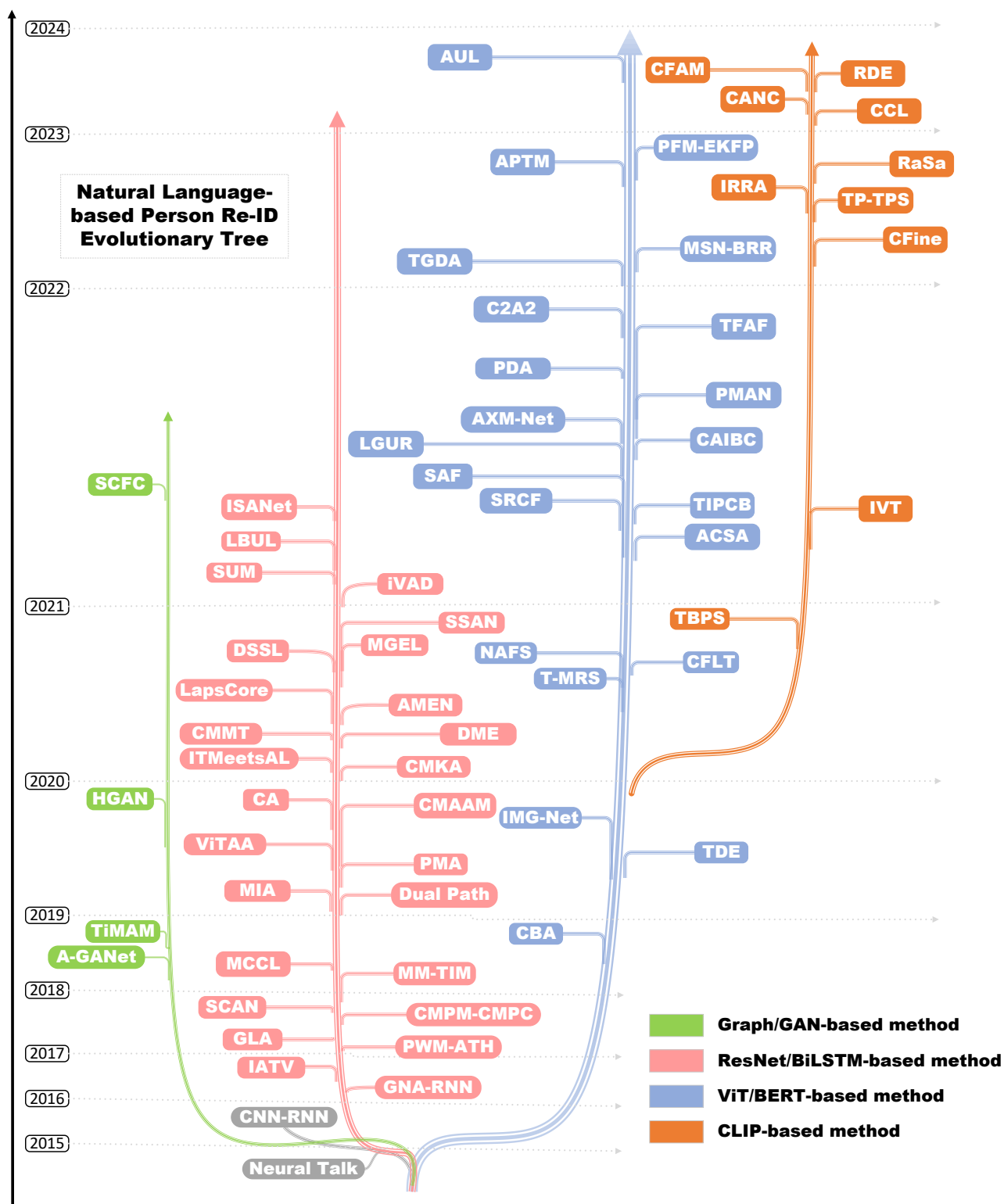
Non-local and contextual attention mechanisms model long-term dependencies and align features based on semantics. Farooq et al. [90] applied non-local attention after calculating interactions between text features to model dependencies between text phrases. Inspired by self-attention, Gao et al. [82] proposed Context Non-Local Attention to enable cross-modal features to align in a coarse-to-fine manner based on semantics, rather than relying on predefined rules. Wang et al. introduced a Part-based Multi-scale Attention Network (PMAN) [79] with attention-based branches in a dual-path feature extraction framework, extracting visual semantic features from different scales to match with text features.

### 3.3.4. Cross-Modal Attention

Lee et al. [91] implemented a novel Stacked Cross Attention to discover complete potential alignments and infer image text similarity using image regions and words in sentences as context. Jiang et al. [40] proposed a new multimodal interaction encoder where textual and visual features are fused through a cross-Modal attention layer and then input into a single transformer block to improve the learning of global image textual representation in the joint embedding space. Chen et al. [93] introduced cross-attention making it possible for the embedding of one modality to be complemented by the aggregated embedding of another modality in the neighborhood, thus avoiding the loss of modality-specific information.

### 3.3.5. Hard Attention

Different from the above soft attention mechanisms, Hard attention select strongly semantically related regions to enhance precise alignment and reduce redundancy. Wang



**Figure 4:** The figure shows a tree development diagram of all existing text-based pedestrian re-recognition methods. In the vertical direction from bottom to top represents the sequence of time development, and in the horizontal direction different trunk colors represent different trunk methods. **Green** represents Graph/GAN-based methods, **pink** represents ResNet/BiLSTM-based methods, **blue** represents ViT/BERT-based methods, **orange** represents CLIP multimodal macromodel-based methods, and **grey** represents text-based pedestrian re-recognition research prior to its emergence, which also appears as a comparison method.

et al. [86] proposed the IMG-Net model, combining intra-modal self-attention and cross-modal hard region attention with a fine-grained model to extract multi-granularity semantic information. Zheng et al. introduced a Hierarchical Gumbel Hard Attention Module [21], using the Gumbel top- $k$  reparameterization algorithm to select semantically related regions for images and corresponding words or phrases. Jing et al. introduced a Cascade Attention Network (CAN) to progressively select from character image and text image similarities [88], specifically involving a similarity-based hard attention to select the similarity scores associated with descriptions from localized similarities.

### 3.3.6. Summary

While attention mechanisms play a crucial role in enhancing cross-modal alignment for fine-grained matching, they often incur higher computational costs due to pairwise inputs, with complexity rising to  $O(MN)$  for  $M$  galleries and  $N$  queries. By categorizing attention mechanisms employed, we can better understand the contributions and approaches of different works.

## 3.4. External Auxiliary

In text-based person Re-ID tasks, a variety of external auxiliary strategies are employed to extract advanced semantic information and enhance cross-modal retrieval performance. These strategies include random masking [94], noise filtering [95], human body keypoints [17], attribute prediction [24], clustering analysis [36], knowledge distillation [96], information theory [97], and color extraction [98].

**Random Masking** Random masking techniques compel models to explore additional useful matching clues, increasing data diversity and improving generalization capabilities. Shu et al. [94] proposed Bidirectional Mask Modeling (BMM), which does not require manual annotation. Random masks are applied to both images and their corresponding textual keywords. Ref. [40] used a static embedding of masked textual tokens as a local fine-grained key to align image and textual contextual representations in the same context.

**Noise Filtering** Reed et al. [95] introduced a method that leverages complementary information from accurate and noisy body parts to update representations. Similarly, Suo et al. proposed a Simple Robust Correlated Filtering (SRCF) framework [99]. This framework employs denoising filters and dictionary filters to construct semantic templates and compute similarities between templates and inputs.

**Pose Information** Pose information and human body keypoints can be leveraged to learn semantic alignments between visual body parts and textual descriptions. Ref. [88] recommended a pose-guided network focusing on the person in the augmented input, which progressively selects key matching cues from the person image and text image similarities. Jing et al. [17] proposed a Pose-guided Multi-granularity Attention Network (PMA), which uses pose information to align visual body parts with textual noun phrases.

**Attribute prediction** Attribute prediction contributes to creating semantically rich embeddings. Wang et al. proposed the Visual-Textual Attribute Alignment (ViTAA) model [24], which utilizes segmentation labels to drive the learning of attribute-aware features from input images. Similarly, aggarwal et al. [100] introduced a text-based person search method that learns attribute-driven spatial information and class-driven spatial information. Attribute prediction, as an auxiliary task, helps to produce semantically preserved embeddings.

**Clustering and Pseudo-Labeling** Clustering methods alleviate intra-class variation by generating pseudo-labels. Zhao et al. [36] proposed a Cross-Modal Mutual Training (CMMT) framework, which includes a mutual pseudo-label refinement module. This module uses clustering results from one modality to refine clustering results in the other modality, constrained by text-image pair relationships. Similarly, Gong et al. proposed an unsupervised cross-modal semantic aligning and neighbor-aware completing (CANC) method [42] based on clustering to establish pedestrian pseudo-labeling and reconstruct text image feature alignment respectively.

**Knowledge Distillation** Knowledge distillation and mutual learning mechanisms enhance the matching capabilities between modalities. Chen et al. [96] introduced knowledge distillation into image and text networks to balance information from both modalities.

**Information Theory** Information theory plays a crucial role in reducing inter-modal and intra-class gaps, thereby enabling effective cross-modal retrieval. Ref. [97] integrated information theory and adversarial learning into an end-to-end framework. Regularization terms based on KL divergence and temperature scaling were introduced to address data imbalance and reduce heterogeneity in cross-modal retrieval.

**Color Dependency** Color significantly affects cross-modal pedestrian retrieval performance. To address the over-reliance on color information, Wang et al. [98] proposed a three-branch CAIBC architecture consisting of RGB, grayscale (GRS), and color (CLR) branches. This architecture effectively mitigates color dependency in cross-modal retrieval tasks.

## 4. Architecture

When features from a specific modality to a common manifold, the feature distribution of other modalities remains imperceptible. This implies that embedding and aligning multimodal features in the common manifold entirely depend on the model's own experience rather than the actual data distribution. In other words, a major challenge in cross-modal Person Re-ID is ensuring that the feature distribution in the common manifold accurately reflects the feature distribution in the original modalities. This necessitates the model to possess sufficient capability to capture and understand the relationships between different modalities, which typically requires abundant data and carefully designed model

structures. Figure 4 shows different branches of the model architecture.

#### 4.1. Convolutional Neural Network

Text-based person Re-ID can be viewed as a multimodal pedestrian retrieval problem, with the primary challenge lying in the extraction of features from both visual and textual data. Due to the capability of neural networks to automatically learn and extract features from pedestrian images, the cumbersome process of manual feature design can be avoided. Most existing efforts consider employing network architectures designed for image classification as the backbone of visual networks. Consequently, convolutional neural networks (CNNs) have been utilized extensively in text-based person Re-ID, primarily leveraging CNNs [8, 34, 90, 87, 101, 88, 102, 103], and their major variants such as VGG-16 [95, 60, 7, 28, 20, 25, 104], MobileNet [105, 100, 32, 27], ResNet-50 [106, 99, 89, 98, 102], ResNet-101 [29, 107, 108, 77, 35], ResNet-152 [97], etc., for extracting crucial information in visual features. These networks are sometimes also employed for text feature extraction [22, 101, 102, 109, 90, 34, 8]. In this section, we review the application of CNNs in text-based person Re-ID, dividing the subsection into two categories: Vision-Based Encoding and Dual-Tower Encoding.

##### 4.1.1. Vision-based Encoding

Vision-based encoding methods primarily use CNNs to extract visual features from pedestrian images. For example, Li et al. [7] first presented the problem of text-based pedestrian retrieval and used a visual sub-network with the same underlying structure as the VGG-16 [110] network to efficiently capture the relationship between words and images. Aggarwal et al. [100] proposed an approach based on learning semantically driven generic embedding of text in images and used the Mobilenet [111] model pre-trained on ImageNet for encoding images. Niu et al. [107] fused source to target orientation and achieved cross-domain adaptation, and their proposed image sentence alignment framework for visual coding uses ResNet-101 [112]. Ref. [108] presented a textual relational embedding approach (TDE) in which the visual coder also uses a ResNet-101 pre-trained network. Jing et al. [113] suggested a visual encoder using a Moment Alignment Network (MAN) of ResNet-50 [112] to solve the cross-modal cross-domain people search task in this paper, in response to the problem of having to manually annotate existing cross-modal data. Gao et al [114] introduced a text-guided denoising and alignment (TGDA) model for image encoders based on ResNet-50 to mitigate information inequality and achieve effective cross-modal matching. Unlike the previous work, Chen et al. [97] utilized ResNet-152 as a visual coder for the model in the process of solving the heterogeneity gap problem for cross-modal retrieval.

##### 4.1.2. Dual-tower Encoding

Dual-tower encoding architectures employ CNNs for both visual and textual feature extraction, creating two parallel encoding networks for images and text. This architecture

aims to align semantic information across modalities in a shared embedding space. Zheng et al. [22] presented an end-to-end Dual-path CNN for the fine-grained problem of image-text matching that may not be satisfied by the pre-trained models, which can extract image and text features from the data more efficiently. Farooq et al. proposed a two-stream deep network framework supervised by cross-entropy loss [101]. The linguistic network is changed to a deep residual network with a similar number of layers as the visual branch, and weights are shared in the last two layers of the network. Ma et al. proposed a novel dual-path CNN with maximal gating block (DCMG) to extract differentiated word embeddings [109], and the proposed framework is based on two ResNet-50 models and makes visual-textual associations more focused on the salient features of both modalities. Wang et al. [98] pre-trained two independent ResNet-50 models on ImageNet and utilized a mutual learning mechanism for multi-branch visual information complementation. Chen et al. [8] suggested that the TIPCB architecture consists of a two-path locally aligned structure, where the visual CNN branch applies a PCB after the backbone network, and the textual CNN branch applies a multi-branch residual network after the pre-trained BERT model. Ref. [90] proposed an architecture AXM-Net based on CNNs that does not rely on external cues for explicit feature alignment. It learns semantically aligned cross-modal feature representations by using CNNs for both modalities.

##### 4.1.3. Summary

CNNs offer significant advantages in localizing perceptual features in pedestrian images and support end-to-end learning. However, they have limitations in handling long-term dependencies and sequential data relationships, which are important for textual data. Additionally, recent trends involve large models such as CLIP have been applied to fine-grained visual tasks, transferring the visual-text contrastive learning paradigm to text-based person Re-ID, to better capture these dependencies and enhance cross-modal alignment [38].

#### 4.2. Recurrent Neural Network

Beyond visual features, text description directly affects the accuracy of text-based person Re-ID [70]. Recurrent Neural Networks (RNN) and their variants have been widely accepted by the research community as neural networks for processing sequential events including text [60, 95, 7, 104].

##### 4.2.1. Early Text Encoding

During the early stages, some work first mapped words to a vector space after word embedding using tools such as NLTK [31, 115, 94, 114] and Stanford CoreNLP [24] to aid in word embedding. Subsequently, inputs to LSTM [20], Bi-LSTM [26, 27, 29, 105, 25, 17, 24, 100, 96, 97, 36, 80, 81, 70, 104], Bi-GRU [91, 107, 106, 31, 35, 115, 116] and other variants were used to try to capture the long-term dependencies of sentences and extract semantic features from text sequences.

Ref. [28] by learning to detect the temporal and spatial features of the segment where the image of the video clip is located using Bi-LSTM, which reduces the variation of the same person in different samples, and facilitates the learning of similarity features. Ding et al. [70] obtained the original part-level texture features by using Bi-LSTM, inferring the part correspondences of words based on the Word Attention Module (WAM) of the word representations, and obtaining the original part-level texture features with reference to the part correspondences of the words. Yan et al. [81] adopted a recursive feature aggregation network based on LSTM to accumulate discriminative features from the first to the deepest node, which effectively mitigates the interference caused by occlusion, background noise and detection failure. Niu et al. [107] focused on solving visual and semantic disparity problems, innovatively integrates cross-domain adaptation of source and target directions, and its text processing uses Bi-GRUs. Wang et al. [106] proposed using Bi-GRUs for textual feature extraction for intermodal reconstruction and intramodal reconstruction paradigm collaborate with each other to embed features correctly into the opposite modal space. Wang et al. [116] suggested using Bi-GRU to learn text features for learning Consistent Cross-Modal Common Manifold (C3M) for text-based person retrieval, which leads to a more reliable cross-modal distribution consensus than blind prediction.

#### 4.2.2. Recent Text Encoding

Recently, with the emergence of large-scale language models such as Bidirectional Transformer Encoder Representation (BERT) [117], the paradigm of pre-training + fine-tuning has triggered a shift in the field of text-based person Re-ID [29]. Pre-trained models are utilized to learn general features of natural language and obtain high-quality word embeddings, which are then fed into RNN variants to capture contextual dependencies [21, 35, 98, 118].

Addressing the challenges stemming from large word variance in the textual domain and accurately measuring the distance between two modal features, BERT, a publicly available large language model for extracting word embeddings, is demonstrated for the first time to be successfully applied to the field of text-to-image matching [29]. Zheng et al. [21] presented a new model using BERT and Bi-GRU employing a hierarchical adaptive matching strategy from three different granularities, word, phrase, and sentence-level for fine-grained matching. To address the lack of large-scale datasets, Han et al. [35] introduced a cross-modal momentum comparison learning framework using CLIP text encoder and Bi-GRU for transfer learning that can transmit useful information in the presence of large domain gaps. Wang et al. [98] borrowed the mechanism of mutual learning and proposed a jointly optimized multi-branch architecture incorporating BERT and Bi-GRU networks to make balanced and effective use of the full range of information. Aiming at the problem that existing works usually ignore the differences in feature granularity between modalities, we propose LGUR, an end-to-end framework for text branching

based on BERT and Bi-LSTM, to provide a uniformly granular feature space for the two modalities to extract diverse and semantically consistent features, thus further improving the ReID performance [118].

#### 4.2.3. Summary

However, there are some limitations of RNNs and their variants for text-based person Re-ID tasks. For instance, it is typically sensitive to the embedding quality of pedestrian text inputs, it is typically less efficient in terms of parallelism, it is unable to capture complex global contextual dependencies, and RNNs can only establish temporal correlations on high-level features, and thus are unable to capture relational clues describing spatially localized details of an image.

### 4.3. Graph Neural Network

For solving the spatial relationship modeling problem of CNN and RNN, text-based person Re-ID using Graph Neural Networks (GNN) is also an early research direction [30, 106, 84]. GNN is a deep learning method based on graph structure for processing unstructured data and its relationships, and is widely used for a variety of data types such as image, text and structured data. In the text-based Person Re-ID, each node in the GNN represents a pedestrian body part and its corresponding phrase description, and the edges between the nodes represent the semantic relationships between different body parts and descriptions, such as verb (wear), noun (female), adjective (yellow), etc. In modeling, first, the image of each person and its corresponding textual description are extracted from the dataset. After that, two Graph Convolution structures are constructed based on auxiliary strategies (e.g., Attention Mechanism, Image Segmentation, etc.). Subsequently, a variant such as Graph Convolutional Networks (GCN) is utilized to learn the relationships between the body nodes in the graph structures and convert these relationships into feature vectors. Finally, the feature vectors of the body structures are asymmetrically aligned across modalities, and the features are fed into a classifier for the person Re-ID.

Several studies have explored GNN-based pedestrian recognition methods. For example, Liu et al. [30] presented a novel Deep Adversarial Graph Attention Convolutional Network (A-GANet) for textual pedestrian retrieval. A-GANet combines pedestrian textual queries with textual and visual scene graphs from gallery images, including object attributes and relationships, from which it learns informative textual and visual representations. Notably, the model also builds a more efficient joint latent feature space between text and vision through adversarial learning, bridging the inter-modal gap and facilitating pedestrian matching. Li et al. [84] proposed a graph convolution-based underlying framework as a feature representation. Similarly, they simulate adversarial attacks on feature extraction induced by textual and image diversity by embedding additional attack nodes in the graph convolution layer, thus enhancing the robustness of the model to textual and visual diversity.

The advantages of GNN in person Re-ID include, on the one hand, the ability to capture semantic relations and model

the semantic relationships between different pedestrian images and textual descriptions, which helps to achieve inferential retrieval and identify pedestrians involved in complex events. On the other hand, GNNs are able to handle textual descriptions of different lengths and structures with higher data flexibility. However, there are some limitations, which may lead to higher computational complexity and memory requirements when GNNs are computing edges of pedestrian body nodes and entity relationships. In addition, GNNs are more susceptible to overfitting than generic network structures.

#### 4.4. Autoencoder

In text-based Person Re-ID tasks, Autoencoder architectures have also been utilized. These models typically map images and textual descriptions into a low-dimensional embedding space, which is then used for person Re-ID [20]. Through this process, pedestrian images and their corresponding textual descriptions are transformed into low-dimensional vector representations, extracting key semantic information to accurately describe the pedestrian's identity [35]. Compared with traditional methods, autoencoders, especially those designed with non-standard encoders, show significant advantages in text-based person Re-ID. Foremost, enhanced semantic representation, which can effectively solve the problem of lack of semantic information in image-text pairs. In Moreover, efficient feature extraction, they can capture key information from text descriptions while compressing the data to reduce redundancy. Finally, improved interpretability, the relationship between image and text descriptions can be visualized in the embedding space through vector relations.

For example, Wang et al. [106] proposed the Adversarial Multi-space Embedding Network (AMEN). This model uses cross-modal and intra-modal reconstruction to embed features into opposing modal spaces while building a robust shared embedding space. Based on an encoder-decoder framework, AMEN effectively aligns and matches embeddings across modalities. Similarly, Ref. [104] introduced a method called Virtual Attribute Decoupling (iVAD) to improve embedding learning. This model employs an encoder-decoder structure to decompose attribute information from images and text into hidden vectors, generating attribute-related embeddings. A hierarchical feature embedding framework was designed, incorporating the Attribute-Enhanced Feature Embedding (AEFE) module. This module integrates attribute-related embeddings into the learned image-text embeddings, enhancing the discriminative power of the features. However, these methods also have limitations. First, training Autoencoder-based models requires a large amount of labeled data. Second, fine-tuning the embedding space dimensions and model parameters is essential to achieve optimal performance.

#### 4.5. Transformer

The Transformer not only has deep modeling capabilities and scalability, but also demonstrates strong modality adaptation. In multimodal tasks, such as Text-based Person

Re-ID, Transformer has become an important technology foundation in this field through the efficient extraction and alignment of textual and visual features. [77, 82, 119, 120, 78, 8, 99, 84, 118, 98, 90, 79, 2, 85, 34, 114, 121, 37, 122]. [35, 38, 39, 40, 41, 42, 43, 44] [123] According to the way the model handles text and visual modalities, the current technological frameworks can be classified into two categories: Dual-stream Encoding and Unified Large Models Era Encoding.

##### 4.5.1. Dual-stream Encoding

The Dual-stream Encoding framework uses independent textual and visual encoders to perform feature extraction on data from both modalities. Each modal encoder (e.g., BERT, Swin Transformer) focuses on its specific data type, e.g., textual modality uses Transformer [94] and BERT [8, 99, 84, 118, 98, 90, 79, 2, 85, 34, 114, 121, 122, 45], etc. to extract contextual features of text. Visual modality uses Vision Transformer [124, 121], Swin Transformer [78, 37, 45], pyramid ViT [85] etc. to extract multi-scale features of person images. Afterwards, the features of both modalities are projected into the same space for matching by a specific alignment mechanism (e.g., Contrastive Learning).

Some work has focused on the use of transformers to solve problems in scale alignment and granularity representation. Li et al. [119] proposed the Semantically-Aligned Feature aggregation network (SAF) based on Vision Transformer [125], which adaptively aggregates unitary features with the same semantics into distinct partial features. Ji et al. [78] introduced an Asymmetric Cross-Scale Alignment (ACSA) method using Swin Transformer to globally align images and text, then dynamically align cross-modal entities, addressing hypothetical scale alignment issues. Shao et al. presented the Learning Granularity-Unified Representations (LGUR) framework [118], a Dei-Transformer-based model that maps visual and textual features into a granularity-unified feature space.

To mitigate alignment flaws from weakly supervised learning, Ref. [85] proposed a joint tokenization and feature alignment framework based on Pyramid Vision Transformer, progressively reducing intermodal and intraclass gaps. Focusing on inter-modal noise, Li et al. [121] introduced the Multi-Granularity Separation Network with Bidirectional Refinement Regularization (MSN-BRR), utilizing only BERT encoders for text encoding. Yang et al. [37] developed a Swin Transformer-based framework called Joint Attribute Prompt Learning and Text Matching Learning (APTML), which enhances representation learning by explicitly incorporating attribute recognition. Ref. [122] proposed a Vision Transformer-based progressive feature mining and external knowledge-assisted feature purification method, enabling the model to extract discriminative features from neglected information. Li et al. [45] addressed the challenge of inherent heterogeneous modal gaps amid significant intra and inter class variation by proposing the Adaptive Uncertainty Based Learning (AUL) framework based on Swin Transformer. Distinct from methods using separate

networks, Shu et al. introduced the Implicit Visual Text (IVT) framework [94], where textual and image feature extraction share parameters within a Transformer backbone network. This approach, optimized using cross-modal data, facilitates learning a shared spatial mapping.

#### 4.5.2. Unified Large Models Era Encoding

With the rise of large-scale pre-trained multimodal models, Unified Large Models provides a unified Transformer-based framework for deep inter-modal feature fusion through shared parameters or joint pre-training, and demonstrates significant advantages in generalization capability and performance [35, 38]. Specifically, a general cross-modal representation is learned by pre-training using a large-scale image-text pair dataset (e.g., 400 million pairs data from CLIP [126]). Since the fine-grained features required for text-based person Re-ID have been learned during the pre-training process, upon which the use of Unified Large Models can be quickly fine-tuned to adapt to downstream tasks [39, 40, 41, 42, 43, 44].

Yan et al [38] firstly proposed the use of CLIP visual coder-driven fine-grained information mining framework (CFine) to fully utilize the powerful knowledge of CLIP for TIREID for effective fine-grained information mining. And Ref. [39] considered by using only pre-trained visual coders and ignoring the corresponding textual representations, breaking the important modal alignment learned from large-scale pre-training. TPS models that fully utilize two CLIP pre-training models in TPS tasks are explored. The IRRRA method [40] addressed the intra-modal information distortion issue through implicit relational inference and alignment without prior supervised learning of local visual-textual relations. Zuo et al. presented CFAM [44], a CLIP-based architecture for ultra-fine-grained text retrieval, which achieves fine-grained mining by employing a shared cross-modal granularity decoder and a hard-negative matching mechanism.

In response to data privacy issues and data noise limitations in real-world open scenarios, Du et al. [41] suggested Incomplete Textual Image Pedestrian Re-recognition (iTIREID), which utilizes clustering features to supplement missing modal features, facilitating effective training with incomplete data. Gong et al. [42], meanwhile, explored increasing interclass distance to suppress intraclass variation and introduced nearest-neighbor consistent complementation to recover high-quality complementary features for robust text-image pedestrian Re-ID (RTIREID). Qin et al. proposed a new robust double embedding approach (RDE) to learn robust visual semantic associations preventing the model from collapsing under data noise [43], and also focusing on hard-negative samples for promising performance. All three use CLIP-ViT and CLIP-Xformer as image and text encoders.

#### 4.5.3. Summary

Transformer has demonstrated its strong cross-modal adaptation capability and feature extraction advantages in

Text-Based Re-ID. There are also some challenges, such as the ability of cross-modal spatio-temporal alignment, fine-grained feature mining, and complex semantic reasoning, as well as the need to improve the efficiency, robustness, and privacy preservation level of the model. By integrating the innovation of large-scale pre-trained models and lightweight design, Transformer is expected to further promote Text-Based Re-ID towards accuracy and ubiquity.

## 5. Optimization

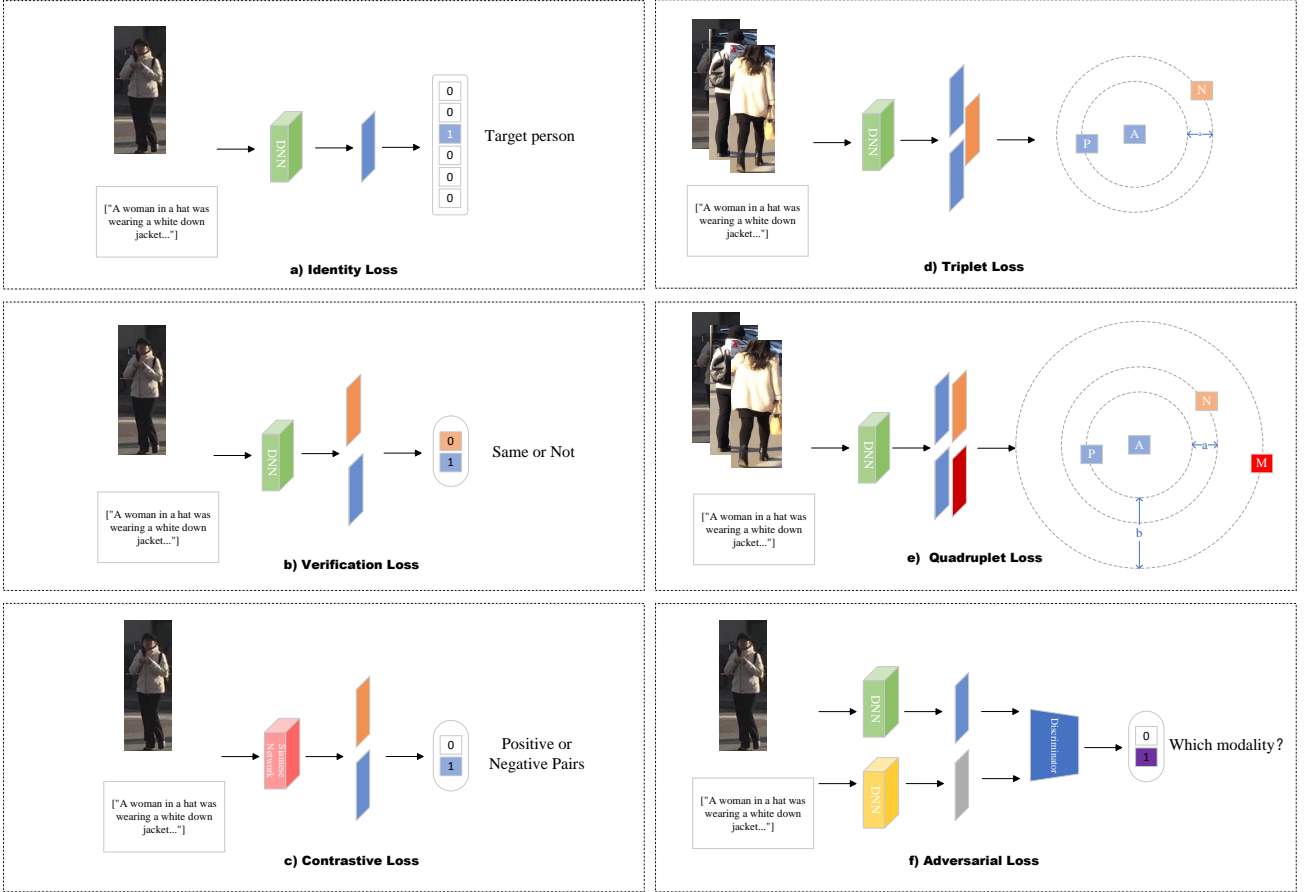
Deep metric learning is crucial for optimization in cross-modal person Re-ID [127]. For instance, it enhances the discriminative power of feature representation, addresses a multitude of noise and variations such as diverse lighting conditions, poses, and occlusions, adapts the distance metric accordingly, embeds more semantically meaningful spatial structures, and leverages domain adaptation to tackle the challenges posed by multiple viewing angles. Of particular significance is the optimization of loss functions, thereby improving the matching performance in cross-modal person Re-ID tasks [128]. This approach effectively addresses a spectrum of challenges, elevating the network's robustness and generalization capacity, consequently achieving more accurate person matching across distinct camera scenes and conditions. This study places emphasis on prevalent loss functions used in text-based person Re-ID tasks, including identity loss, verification loss, contrastive loss, triplet loss, quadruplet loss, and adversarial loss, as shown in Figure 5.

### 5.1. Identity Loss

Zheng et al. [129] introduced the ID Discriminative Embedding (IDE) network, which treats each pedestrian as a separate category and employs the pedestrian's ID as a classification label to train a deep neural network. This allows the network to learn the capability of predicting the identity given a pedestrian image. The classification loss is typically computed using the cross-entropy loss function. During training, the classifier optimizes its weight parameters by minimizing the classification loss. A fully connected layer (FC) for classification is appended to the end of the network, and the softmax activation function maps the features to a probability space representing identity labels [103]. The cross-entropy loss for the multi-class person Re-ID task can be expressed as:

$$L_{ID} = -\frac{1}{N} \sum_{i=1}^N y_i \log(p_i). \quad (4)$$

Here,  $x_i$  denotes the input image,  $y_i$  represents the corresponding label,  $p_i$  is the probability that  $x_i$  is recognized as  $y_i$  after the softmax function, and  $N$  indicates the batch size. Identity loss is widely adopted in person Re-ID [98, 102, 125, 29, 106, 99, 101, 130, 115], due to its automatic hard sample mining and ease of training. However, as the labeled data increases significantly, issues related to decreased efficiency in classifier training might



**Figure 5:** Overview of key alignment loss methods for text-based person re-identification. These include: (a) Identity Loss, which aligns textual descriptions with identity labels to ensure correct identification; (b) Verification Loss, which verifies whether a text-image pair belongs to the same identity; (c) Contrastive Loss, which reduces intra-class distances while increasing inter-class distances for paired samples; (d) Triplet Loss, which optimizes the relative distances among anchor, positive, and negative samples; (e) Quadruplet Loss, which further refines intra-class compactness and inter-class separability by introducing additional constraints; and (f) Adversarial Loss, which employs adversarial learning to produce modality-agnostic feature representations. These methods form the basis of most contemporary alignment strategies through their variants or combinations.

arise, especially when dealing with datasets containing a large number of similar pedestrians. In cross-modal person Re-ID, the classification loss is often used in conjunction with other loss functions [36, 35], such as contrastive loss and triplet loss [21, 20, 81, 98], to enhance performance of the model.

## 5.2. Verification Loss

The verification loss is employed to quantify the similarity between two pedestrian images [131, 22]. In person Re-ID, it is often necessary to compare two images to determine if they belong to the same individual. The verification loss aims to achieve this objective by comparing the feature vectors of two images to compute their similarity [127]. Specifically, the verification loss is calculated by inputting the feature vectors of two images into a similarity metric function. This similarity metric function can be Euclidean distance, cosine similarity, or other similar measures. If two images belong to the same person, their feature vectors should exhibit high similarity, resulting in a high similarity

score [72]. Thus, the verification loss is minimized to enhance the similarity between two images of the same person. The specific formula is as follows:

$$L_{veri} = -\frac{1}{N} \sum_{i=1}^N y_i \log(p_i) + (1 - y_i) \log(1 - p_i). \quad (5)$$

Here,  $N$  is the number of sample pairs,  $y_i$  is a label indicating whether the sample pair  $x_i$  and  $x_j$  belongs to the same category (1 denotes the same category, 0 denotes different categories), and the softmax function is utilized to calculate the probability  $p_i$  of having the same label for both samples. Due to the necessity of pairwise computation, the verification loss tends to have lower efficiency in Re-ID tasks. Therefore, it is often used in combination with other losses, such as identity loss, to leverage the advantages of both methods, thereby enhancing recognition speed and accuracy.

### 5.3. Contrastive Loss

Contrastive loss was originally introduced by [132] in the context of face verification tasks, aiming to train neural networks to determine whether two inputs belong to the same facial identity. In the text-based person Re-ID [24, 77, 35, 123, 37], the principle of encouraging similar identity feature representations to be closer in the feature space while dispersing feature representations of different identities is leveraged for model training within Siamese networks [133]. The Contrastive Loss not only considers the distances between images of the same identity but also takes into account the distances between images of different identities. The function can be expressed as follows:

$$L_{con} = \frac{1}{2N} \sum_{n=1}^N y_n D_n^2 + (1 - y_n) \max(\text{margin} - D_n, 0)^2. \quad (6)$$

Here,  $N$  is the number of sample pairs,  $y_n$  is a label indicating whether sample pairs  $x_i$  and  $x_j$  belong to the same class (1 for the same class, 0 for different classes),  $D_n$  represents the distance between feature representations (Euclidean distance, cosine distance, etc.), and  $\text{margin}$  is a threshold indicating that when sample pairs belong to different classes, their distance should be greater than this value. For sample pairs belonging to the same class, the loss is the square of the distance  $D_n$ , while for sample pairs belonging to different classes, the loss is the square of the difference between the distance and the threshold value  $\text{margin}$ . This framework aims to bring feature representations of sample pairs from the same class closer together and push feature representations of sample pairs from different classes further apart. By adjusting the value of  $\text{margin}$  and the weight of the loss, the distance range between sample pairs can be controlled, influencing the learning of feature representations [134]. Currently, CPM-CPM Loss can be regarded as a variant of contrastive loss, which was first proposed by [127] and has since been widely used by text-based person Re-ID [30, 108, 36, 82, 120, 78, 8, 85, 79, 94, 84, 34, 38, 121, 41].

### 5.4. Triplet Loss

The triplet loss is a commonly used loss function in text-based person Re-ID, designed to learn discriminative feature representations by comparing the distances between three samples: a text description (anchor), a corresponding image (positive), and an unrelated image (negative) [80, 98, 135, 136, 106, 90, 81, 115, 23, 137, 38, 43]. The fundamental idea is to align cross-modal feature representations by minimizing the distance between the text and its corresponding image while maximizing the distance between the text and an unrelated image [88, 103].

Specifically, given an anchor text, a positive image, and a negative image, the triplet loss aims to simultaneously minimize the distance between the anchor text and the positive image and maximize the distance between the anchor text and the negative image [21, 105]. The formal definition is as follows:

$$L_{triplet} = \sum_{n=1}^N \max(\|f_{\text{text}}(t_n^a) - f_{\text{img}}(i_n^p)\|_2^2 - \|f_{\text{text}}(t_n^a) - f_{\text{img}}(i_n^n)\|_2^2 + \text{margin}, 0), \quad (7)$$

Here,  $f_{\text{text}}$  is a function that maps a text description  $t$  to a feature vector, while  $f_{\text{img}}$  maps an image  $i$  to a feature vector in the same feature space.  $t_n^a$ ,  $i_n^p$ , and  $i_n^n$  represent the anchor text, the positive image corresponding to the anchor text, and the negative image unrelated to the anchor text of the  $n$ -th instance, respectively.  $\|\cdot\|_2$  denotes the Euclidean distance between two feature vectors. The term  $\max(\cdot)$  ensures that the loss is zero if the margin condition is satisfied. The hyperparameter  $\text{margin}$  defines the minimum required distance difference between the anchor-positive pair and the anchor-negative pair, encouraging a clear separation of feature representations for different identities.

It is noteworthy that in practice, triplet loss has encountered certain issues, such as slow learning and effective sample selection. To address these issues, several improved variants have been proposed. For example, Ref. [105] introduced the Mutually Connected Classification Loss (MCCL), which enhances cross-modal alignment by incorporating identity-level information. This method not only embeds identity information into both text and image modalities but also encourages similar classification probabilities for instances of the same identity across modalities. These enhancements have proven effective in improving performance for text-based cross-modal person Re-ID.

### 5.5. Quadruplet Loss

The quadruplet loss extends the triplet loss by introducing a fourth sample, specifically designed for person Re-ID tasks. Unlike the triplet loss, which aligns features between a text description (anchor), a corresponding image (positive), and an unrelated image (negative), the quadruplet loss introduces an additional negative image to improve stability and discriminative performance [138]. The added constraint ensures that the model not only aligns features between matched pairs but also enforces larger separations between unrelated samples.

Specifically, the quadruplet loss compares embedding distances between an anchor text, a positive image, a first negative image, and a second negative image. The objective is to ensure that the distance between the anchor text and the positive image is smaller than the distance between the anchor text and the first negative image by a margin  $\alpha$ , and that the distance between the first and second negative images is larger than the distance between the anchor text and the first negative image by a margin  $\beta$ . The formal definition is as follows:

$$L_{quad} = \sum_{n=1}^N \max(\|f_{\text{text}}(t_n^a) - f_{\text{img}}(i_n^p)\|_2^2 - \|f_{\text{text}}(t_n^a) - f_{\text{img}}(i_n^{n1})\|_2^2 + \alpha, 0) + \max(\|f_{\text{text}}(t_n^a) - f_{\text{img}}(i_n^{n1})\|_2^2 - \|f_{\text{img}}(i_n^{n1}) - f_{\text{img}}(i_n^{n2})\|_2^2 + \beta, 0). \quad (8)$$

Here,  $f_{\text{text}}$  and  $f_{\text{img}}$  are the embedding functions for text and image inputs, respectively, mapping them into a shared feature space.  $t_n^a$ ,  $i_n^p$ ,  $i_n^{n1}$ , and  $i_n^{n2}$  represent the anchor text, the positive image, the first negative image, and the second negative image of the  $n$ -th instance, respectively.  $\|\cdot\|_2$  denotes the Euclidean distance between two feature vectors. The hyperparameters  $\alpha$  and  $\beta$  define the required margins to separate distances between positive and negative pairs. The  $\max(\cdot)$  function ensures that the loss is zero when the margin conditions are satisfied.

By introducing an additional negative image, the Quadruplet Loss reduces within-class variance and improves inter-class separability for cross-modal alignment. However, it also increases computational complexity due to the need for more sample pairs. Efficient sampling strategies are critical to balance computational cost and performance gains in practical applications.

### 5.6. Adversarial Loss

The adversarial loss primarily leverages adversarial training to learn more robust pedestrian feature representations [84]. Its fundamental concept involves mapping pedestrian feature embeddings from the training data into a similarity space, where the distribution of generated image features is made closer to that of real image features, enhancing the robustness of the network in handling multimodal information [29]. Specifically, adversarial loss employs adversarial samples during training, where the model seeks to minimize the adversarial loss. The computation of the adversarial loss involves introducing perturbations to the model input, making the input pedestrian image appear visually similar to the original image while introducing significant disparities in the embedding space. These perturbations are generated by a generator network, which attempts to produce images similar in appearance to the input image but with substantial deviations in the embedding space. The discriminator seeks to maximize the adversarial loss to effectively discriminate between these two types of features. The general formula for adversarial loss is as follows:

$$L_{adv} = \mathbb{E}_{x \sim p_{data}(x)} [\log D(x)] + \mathbb{E}_{z \sim p_z(z)} [\log(1 - D(G(z)))]. \quad (9)$$

Here,  $G$  is the generator,  $D$  is the discriminator,  $x$  represents real image features,  $z$  represents generated image features,  $\log D(x)$  denotes the logarithm of the discriminator's predicted probability for real image features, and  $\log(1 -$

$D(G(z)))$  denotes the logarithm of the discriminator's predicted probability for generated image features. To achieve modality invariance, adversarial loss has been adapted for text and image feature embeddings. The modality discriminator classifies the modality labels (text or image) of the input samples. A well-trained feature network aims to confuse the discriminator by aligning text and image embeddings, ensuring that the discriminator cannot reliably distinguish between modalities [30]. This modality-invariant representation facilitates robust cross-modal retrieval. Although adversarial loss has been relatively recently applied to person Re-ID tasks, it has shown notable efficacy in aligning modality-invariant, semantically discriminative features, as well as improving resistance to adversarial attacks [30, 106].

## 6. Challenge

In this paper, we first give an overview of text-based person Re-ID in two parts: attribute-based and natural language-based person Re-ID. Secondly, the datasets related to these two parts are summarized and described respectively. Moreover, we analyze the commonly used technical strategies in text-based person Re-ID. In addition, an overview of the deep learning architectures frequently employed in text-based person Re-ID is presented. Further, we detail the alignment optimization methods commonly used in this field. Finally, we analyze the challenges and possible future directions of text-based person Re-ID from the perspectives of *architecture challenges*, *closed-world data*, *scene shift*, *noise issues*, *inefficient computation*, and *privacy protection*.

Although text-based person Re-ID technology offers numerous advantages in fields such as video surveillance, security, social media, and virtual reality [82], it still faces various challenges. The most critical issues include the ambiguity and granularity of natural language. Identical descriptions may be incomplete or contain ambiguities, potentially leading to incorrect identity matches. Architectural challenges arise when designing models that can effectively integrate and process multimodal data, requiring complex networks that may be difficult to optimize.

On the other hand, photographs available in real-world galleries pertaining to pedestrians often lack high resolution, ample lighting, and freedom from obstruction [101]. Scene shift and noise issues, such as varying environmental conditions and image quality degradation, further complicate the Re-ID process by affecting the reliability of feature extraction. Additionally, the reliance on closed-world data—datasets that do not represent the full diversity of real-world scenarios—limits the generalizability of models to new, unseen environments.

Zhang et al. [139] demonstrated recently that gait information may be the basis for general person Re-ID, and that its features may be more robust compared to the highly homogeneous pedestrian appearance feature signature. However, existing gait recognition methods are not able to meet the requirements of less controlled scenarios, e.g., most studies are based on well-controlled environments such as

**Table 2**

The table demonstrates the results of person Re-ID with and without fine-tuning the stable diffusion model. Tests are performed on three datasets, CUHK-PEDES, RSTPReID and ICFG-PEDES, respectively, using different visual representation backbone networks.

| Stable Diffusion | Method          | CUHK-PEDES    |               |               | RSTPReID      |               |               | ICFG-PEDES    |               |               |
|------------------|-----------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|
|                  |                 | R@1           | R@5           | R@10          | R@1           | R@5           | R@10          | R@1           | R@5           | R@10          |
| Non-Finetuned    | DenseNet        | 0.0396        | 0.1072        | 0.1944        | 0.0275        | 0.0980        | 0.1882        | 0.0077        | 0.0210        | 0.0323        |
|                  | EfficientNet    | 0.0584        | 0.1233        | 0.2216        | 0.0392        | 0.1255        | 0.2000        | 0.0087        | 0.0209        | 0.0326        |
|                  | SwinTransformer | 0.0630        | 0.1519        | 0.2460        | 0.0471        | 0.1294        | 0.2235        | 0.0084        | 0.0215        | 0.0326        |
|                  | ResNet          | 0.0674        | 0.1794        | 0.2578        | 0.0491        | 0.1353        | 0.2348        | 0.0103        | 0.0287        | 0.0392        |
| Finetuned        | DenseNet        | 0.1101        | 0.2179        | 0.3298        | 0.0745        | 0.2196        | 0.3176        | 0.0328        | 0.0783        | 0.1278        |
|                  | EfficientNet    | 0.1498        | 0.2619        | 0.4305        | 0.1103        | 0.2784        | 0.3686        | 0.0651        | 0.1254        | 0.1701        |
|                  | SwinTransformer | <b>0.1622</b> | <b>0.3794</b> | 0.4675        | 0.1098        | 0.2667        | 0.3804        | 0.0632        | 0.1302        | 0.2015        |
|                  | ResNet          | 0.1586        | 0.3648        | <b>0.4827</b> | <b>0.1176</b> | <b>0.2824</b> | <b>0.4118</b> | <b>0.0749</b> | <b>0.1306</b> | <b>0.2191</b> |

narrow corridors. If one wants to improve the efficiency of person Re-ID by combining gait features with appearance features, it may be necessary to obtain more coherent and high-quality gait data while the appearance is recognizable. This introduces inefficiencies in computation, as processing additional modalities requires more computational resources, and raises privacy protection concerns due to the sensitive nature of biometric data.

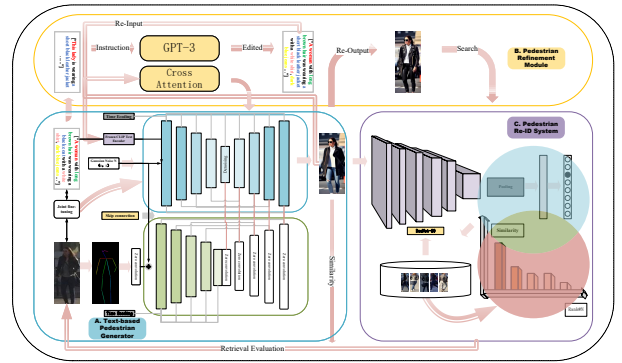
Therefore, combining textual descriptions with image or video information requires more efficient and robust cross-modal fusion techniques that enable the system to better understand and match entities. Addressing architectural challenges and computational inefficiencies may involve utilizing edge computing, and additionally pre-trained large-scale multimodal models may be a reasonable direction to achieve better encoding performance. In addition, text-based person Re-ID requires more high-quality data to enable the model to learn the relationship between various descriptions and identities [37]. However, the collection of such data must be balanced with privacy protection regulations to ensure that personal information is handled securely.

Notably, the recent popularity of text-to-image generation models has significant potential as a new approach to person Re-ID, as these models generate images for retrieval based on textual descriptions. Given that real-world cameras do not always capture images from a library of available queries, it makes sense to customize the generation of corresponding target entities based on witness text. In addition, the use of pedestrian synthetic data for Re-ID largely addresses issues regarding pedestrian privacy protection. However, challenges exist due to scene shifting and noise issues in generating images, as well as the immature controllability of text-to-image generation models. Despite these challenges, we remain optimistic in this direction.

## 7. Foresight

### 7.1. A Novel Baseline

Since the current text-based person Re-ID tasks being based on disparate datasets, specialized models are separately trained on large samples. However, in the real world,



**Figure 6:** Text-Based Pedestrian Image Generation Guided Re-ID baseline architecture. The entire architecture consists of three modules: A) Text-based pedestrian generator, B) Pedestrian refinement module, and C) Pedestrian re-recognition system. The entire architectural flow can be viewed starting with module A at the bottom left, preferably in a clockwise direction. Note that there may not be some kind of very clear boundary between modules A and B, and that there is interaction between them in the pedestrian generation process.

the lack of a uniform style (resolution, size, angle, brightness) in person Re-ID data is often due to the constraints imposed by different specifications of data collection devices. Retrieving arbitrary pedestrian targets in an open-world scenario becomes quite challenging for dedicated models. There is an urgent need for a new approach that enables zero-shot retrieval of targets from any gallery. On the other hand, text-to-image retrieval fundamentally relies on comparing cross-modal data distributions from small datasets through contrastive learning. If the gallery itself lacks target pedestrians, this results in ineffective retrieval, contradicting the concept of open-world pedestrian recognition. Inspired by the controllable generation retrieval in text-to-image for pedestrians, we propose a novel baseline model for text-based person Re-ID, which involves re-retrieving target images from text to image.

As illustrated in Figure 6, the entire text-based pedestrian image generation guided Re-ID baseline architecture is demonstrated. The initial step involves encoding natural

language descriptions of pedestrians and utilizing a diffusion model to generate corresponding pedestrian images. Subsequently, these generated pedestrian images are employed in the task of person Re-ID. Given that current text-to-image generation models primarily focus on coarse-grained overall image generation tasks, emphasizing inter-class coordination among different objects rather than concentrating on intra-class distinctiveness in human images, and while simultaneously considering authenticity, a strategy based on human body key-point assistance is adopted. This strategy involves extracting features from human body key-points to assist in controlling the pedestrian generation process, ensuring compliance with the requirement for the presence of the majority of the human body in the person Re-ID task. The generated pedestrian images serve as queries for retrieval and are input into the backbone network for visual feature extraction. To enhance retrieval accuracy, emphasis is placed on restoring features with the highest relevance to the original image. For instance, in comparison to generated images, sensitivity to channel attention may be more pronounced in original images. Various attention mechanisms are employed to assess their impact on retrieval effectiveness. However, due to the coarser granularity of linguistic descriptions compared to images, the generation of pedestrian images may not be sufficiently detailed in terms of the style and material of the pedestrian's clothing. For example, the textual description may be as broad as "black coat", but the image may be as specific as "short black leather jacket", which results in feature learning inaccuracy. In order to improve the accuracy of localized feature generation for pedestrians in a realistic open set, an approach using cross-attention based image editing is recommended. After using natural language instructions, the local features of pedestrians are controlled to change, and the witness's memory of the local features for the target person is refined and specialized. Finally, the results of pedestrian retrieval are obtained through different optimization strategies.

## 7.2. Experiment & Results

**For A) Text-based pedestrian generation module,** We employed the Frozen CLIP Text Encoder [140] to encode pedestrian descriptions, followed by inputting them into stable diffusion [141] to generate pedestrian images. The denoising module utilized the cross-attention U-Net [142] backbone, which proves beneficial for the generation of high-quality pedestrian images. For pedestrian pose assistance, we adopt ControlNet [143] for Pedestrian integrity control, which ensures that the generated pedestrian image has full body. Specifically, OpenPose [144] was employed for detecting key points on the pedestrian body. Subsequently, pose features were extracted through skip connections and zero convolutions, facilitating their incorporation into pedestrian generation.

**In the part of B) pedestrian local feature refinement,** we input the local image editing instructions and then use GPT-3 [145] to rewrite and replace the pedestrian image

generation instructions. Re-inputting the rewritten text cue and the pedestrian image before rewriting into the stable diffusion model, the visual and textual features of pedestrians are fused using the cross-attention layer, and the image can be edited by editing the text only to realize the mapping from the source image to the target image [146]. As a comparative analysis, we also attempted fine-tuning the baseline model using a subset of text-image pairs data based on text-driven pedestrian Re-ID. In the retrieval module, the generated pedestrian images were employed as queries.

**In the C) pedestrian Re-ID system,** ResNet-50 [112] served as the visual backbone, incorporating self-attention mechanisms. The Part-based Convolutional Baseline (PCB) [147] strategy was employed for segmenting human bodies to extract feature maps, and random horizontal image flipping was applied for data augmentation. The training process involved the use of Instance Batch Normalization (IBN) [148] strategy, with a batch size set to 32, dropout rate at 0.5, erasing probability at 0.2, ins-gamma set to 32, learning rate at 0.05, and weight decay at 0.0005. Furthermore, the training amalgamated ID loss, Contrastive loss, and triplet loss as the loss functions. The optimization utilized SGD optimizer with momentum, and the model was trained for 60 epochs.

As depicted in the Table 2, we evaluated the proposed Text-Based Pedestrian Image Generation Guided Re-ID (TBPGR) baseline architecture on three widely used text-based person Re-ID datasets (CUHK-PEDES, RSTPReid, ICFG-PEDES). Specifically, retrieval of pedestrian images was generated using the generation modules with and without fine-tuned Stable Diffusion, respectively. We report the performance of visual representation modules DenseNet, EfficientNet Swin-Transformer and ResNet on the task for each dataset and each condition of with and without fine tuning. ResNet generally shows better performance. After the image generation in the intermediate step, we qualitatively analyzed the content of the images. It was found that the quality of the pedestrian images generated before fine-tuning was higher, and the quality of the images generated after fine-tuning became lower, but the retrieval efficiency was slightly improved, conjecturing that the quality of the pedestrian image text data pairs had some effect on the data distribution. Besides, since the CUHK-PEDES dataset retrieval effect is better than other datasets, it is hypothesized that it is because the generation of pedestrian images seldom involves the factors such as lens occlusion and illumination changes, which have some implications on the pedestrian retrieval performance, as shown in Figure 7.

Although our baseline model (TBPGR) is still not powerful enough in comparison to traditional training approaches based on implicit spatial alignment of text and images, we believe that such an open-retrieval paradigm is far more valuable in future open-world text-based person Re-ID scenarios than just pursuing the accuracy of the present moment. In future work, we could develop more fine-grained pedestrian detail-controllable generative retrieval models based on this



**Figure 7:** Generated images of text-based pedestrians without and with fine-tuning are shown, along with their retrieval Rank@10 results. The green numbers shown at the top of the image represent a positive retrieval and the red color represents a negative retrieval.

baseline architecture to improve the accuracy of text-based pedestrian generation-guided re-identification tasks.

## Acknowledgments

The authors would like to thank all the anonymous reviewers and associate editors for their valuable comments to improve the paper. The authors would also like to thank the authors of the three publicly available datasets.

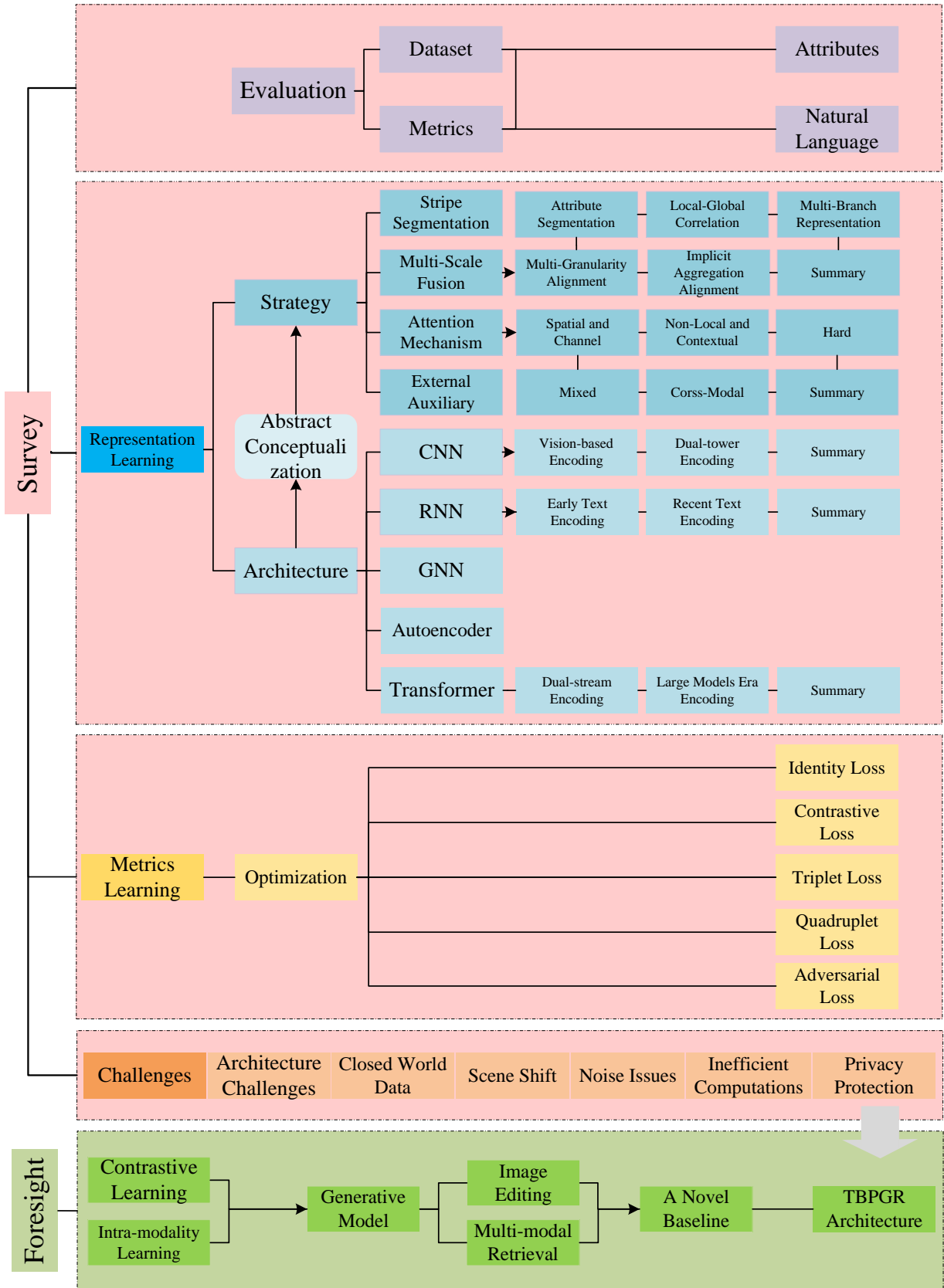
## APPENDIX

This appendix provides supplementary tables and figures that offer detailed information to further support the findings discussed in the main paper, as shown in Figure 8 and Table 4.

**Table 3**

A comparison of surveys in person Re-ID, based on chronological order, highlighting the authors, publication details, and key contributions.

| Author               | Publication  | Ref.    | Contribution  |
|----------------------|--|---------|---|
| Bedagkar-Gala et al. | A survey of approaches and trends in person re-identification                        | IVC14   | Introduced person Re-ID challenges up to 2014 and summarized mainstream solutions.  |
| Zheng et al.         | Person Re-identification: Past, Present and Future                                   | arXiv16 | Discussed past developments in pedestrian retrieval and image classification, investigating image/video-based systems and methods.            |
| Chahar et al.        | Deep convolutional neural network based approaches for person re-identification      | PRMI17  | Presented pedestrian Re-ID methods based on CNNs, highlighting problems and future directions.  |
| Wang et al.          | Survey on person re-identification based on deep learning                            | CAAI18  | Surveyed deep learning-based pedestrian Re-ID approaches, suggesting future paths.  |
| Karanam et al.       | Systematic evaluation and benchmark for person re-identification                     | TPAMI19 | Reviewed and evaluated single- and multi-shot Re-ID algorithms, focusing on feature extraction and metric learning.                           |
| Wu et al.            | Deep learning-based methods for person re-identification: A comprehensive review     | NC19    | Summarized six deep learning methods for Re-ID, including recognition, verification, metrics, parts, video, and data augmentation.            |
| Mathur et al.        | A Brief Survey of Deep Learning Techniques for Person Re-identification              | TCSVT20 | First to describe disparities between closed and open-world scenarios, highlighting open-world Re-ID limitations.                             |
| Islam                | Person search: A survey of recent works  | IVC20   | Covered feature representation learning and deep metric learning with novel loss functions.   |
| Lavi et al.          | Survey on Reliable Deep Learning-Based Person Re-Identification Models               | arXiv20 | Investigated SOTA deep models for person Re-ID and the limitations of these models.   |
| Wang et al.          | Beyond Intra-modality: Heterogeneous Person Re-identification                        | IJCAI20 | Categorized cross-modal Re-ID applications, including sketch, text, low resolution (LR), and infrared (IR).                                   |
| Zou et al.           | Person re-identification based on metric learning: a survey                          | MTA21   | Summarized progress in metric learning-based Re-ID methods.   |
| Ye et al.            | Deep Learning for Person Re-identification: A Survey                                 | TPAMI21 | Reviewed closed/open-world Re-ID, covering feature learning, metric learning, and ranking optimization, and presented his outlook.            |
| Yang et al.          | Survey on Unsupervised Techniques for Person Re-Identification                       | CDS21   | Surveyed SOTA unsupervised approaches in person Re-ID.  |
| Yaghoubi et al.      | SSS-PR: A short survey of surveys in person re-identification                        | PRL21   | Proposed a multi-dimensional taxonomy of Re-ID studies based on diverse viewpoints.   |
| Wang et al.          | Cross-Domain Person Re-identification: A Review                                      | AIC21   | Investigated datasets and compared cross-domain Re-ID methods.  |
| Ming et al.          | Deep learning-based person re-identification: A survey                               | IVC22   | Suggested categorizing deep learning Re-ID methods into metric learning, local feature learning, adversarial learning, and sequence learning. |
| Peng et al.          | Deep Learning-Based Occluded Person Re-Identification: A Survey                      | TMCC22  | Provided a survey on deep learning methods for occluded Re-ID.  |
| Singh et al.         | A comprehensive survey on person re-identification approaches                        | MTA22   | Discussed image/video Re-ID approaches across temporal, spatial, metric, and automation dimensions.   |
| Huang et al.         | Deep learning for visible-infrared cross-modality person re-identification: A review | IF22    | Offered a classification for SOTA visible-infrared cross-modality Re-ID models.   |
| Zahra et al.         | Person re-identification: A retrospective on open challenges and future trends       | PR23    | Overviewed image/video-based Re-ID in four areas: datasets, architecture, number of papers, and challenges.                                   |



**Figure 8:** The figure shows the overall idea and organization regarding the running of the article, which is divided into two parts: survey and foresight. The survey part elucidates the current research on text-based person re-identification and presents some of the existing challenges. For the current problems, a novel baseline architecture is proposed in the foresight part.

Table 4: The table presents all state-of-the-art methods in text-based person Re-ID to date. It categorizes and summarizes these methods across various dimensions, including method names, publications, bidirectional retrieval capability, feature scale, visual backbone, language backbone, auxiliary strategies, alignment losses, and performance on publicly available benchmark datasets.

| Method           | Publication      | Text-to-Image | Feature     | Visual backbone          | Language backbone | Auxiliary strategy                   | Alignment loss                 | R@1   | CUHK-PEDES R@5 | R@10  | R@1   | ICFG-PEDES R@5 | R@10  | R@1   | R@5   | R@10  |
|------------------|------------------|---------------|-------------|--------------------------|-------------------|--------------------------------------|--------------------------------|-------|----------------|-------|-------|----------------|-------|-------|-------|-------|
| Neural Talk [60] | CVPR15           | TI            | Global      | VGG-16                   | Bi-LSTM           | /                                    | ID Loss                        | 13.66 |                | 41.72 |       |                |       |       |       |       |
| CNN-RNN [95]     | CVPR16           | TI            | Global      | VGG-16                   | Bi-LSTM           | /                                    | ID Loss                        | 8.07  |                | 32.47 |       |                |       |       |       |       |
| GNA-RNN [7]      | CVPR17           | TI            | Global      | VGG-16                   | Bi-LSTM           | Attention                            | ID Loss                        | 19.05 |                | 53.64 |       |                |       |       |       |       |
| IATV [20]        | ICCV17           | TI            | Global      | VGG-16                   | LSTM              | Spatial Attention                    | CMCE loss                      | 25.94 |                | 60.48 |       |                |       |       |       |       |
| PWM-ATH [28]     | WACV18           | TI            | Global      | VGG-16                   | Bi-LSTM           | Attention                            | ID Loss                        | 27.14 | 49.45          | 61.02 |       |                |       |       |       |       |
| GLA [26]         | ECCV18           | TI            | Multi-scale | ResNet-50                | Bi-LSTM           | Parts, Mask, NLTK                    | ID Loss                        | 43.58 | 66.93          | 76.26 |       |                |       |       |       |       |
| CMPM-CMPC [27]   | ECCV18           | IT, TI        | Global      | MobileNet                | Bi-LSTM           | /                                    | CMPM-CMPC Loss                 | 49.37 | 71.69          | 79.27 | 43.51 | 65.44          |       |       |       |       |
| SCAN [91]        | ECCV18           | TI, IT        | Multi-scale | ResNet-101, Faster R-CNN | Bi-GRU            | Regions/Attention                    | Triplet Loss                   | 55.86 | 75.97          | 83.69 | 50.05 | 69.65          |       |       |       |       |
| MCCL [105]       | ICASSP19         | TI            | Global      | MobileNet                | Bi-LSTM           | Text map Attention                   | KL divergence, Triplet Loss    | 50.58 |                | 79.06 |       |                |       |       |       |       |
| A-GANet [30]     | ACM MM19         | TI            | Local       | ResNet50/GAC             | Bi-LSTM/GAC       | Regions, GAN, GCN                    | CMPM loss                      | 53.14 | 74.03          | 81.95 |       |                |       |       |       |       |
| TiMAM [29]       | ICCV19           | IT, TI        | Global      | ResNet-101               | LSTM              | GAN, BERT                            | BCE loss                       | 54.51 | 77.56          | 84.78 |       |                |       |       |       |       |
| CBA [107]        | ICCVW19          | TI            | Global      | ResNet-101               | Bi-GRU            | Attention                            | ID/Triplet Loss                | 57.84 | 78.33          | 85.43 |       |                |       |       |       |       |
| Dual Path [22]   | ACM TOMM20       | TI            | Global      | ResNet-50                | ResNet-50         | Data distribution                    | Instance/Ranking Loss          | 44.40 | 66.26          | 75.07 | 38.99 | 59.44          |       |       |       |       |
| MIA [25]         | IEEE TIP20       | TI            | Multi-scale | VGG-16                   | Bi-LSTM           | Regions, NLTK                        | BCE Loss                       | 53.10 | 75.00          | 82.90 | 46.49 | 67.14          |       |       |       |       |
| GALM PMA [17]    | AAAI20           | TI            | Global      | ResNet-50                | Bi-LSTM           | Pose Estimator                       | ID/Ranking Loss                | 53.81 | 73.54          | 81.23 |       |                |       |       |       |       |
| TDE [108]        | ACM MM20         | TI            | Global      | ResNet-101               | BERT              | Attention                            | CMPM loss                      | 55.25 | 77.46          | 84.56 |       |                |       |       |       |       |
| ViTAA [24]       | ECCV20           | TI            | Multi-scale | ResNet-50                | Bi-LSTM           | Mask R-CNN, Stanford CoreNLP         | Attribute/Contrastive Loss     | 55.97 | 75.84          | 83.52 | 50.98 | 68.79          | 75.78 |       |       |       |
| IMG-Net [86]     | JEI20            | TI            | Multi-scale | ResNet-50                | BERT              | Parts                                | ID/Triplet Loss                | 56.48 | 76.89          | 85.01 |       |                |       | 37.60 | 61.15 | 73.55 |
| CMAAM [100]      | WACV20           | TI            | Multi-scale | MobileNet                | Bi-LSTM           | Attributes, NLTK                     | ID/Triplet/Attribute Loss      | 56.68 | 77.18          | 84.86 |       |                |       |       |       |       |
| HGAN [21]        | ACM MM20         | IT, TI        | Multi-scale | ResNet-50                | BERT              | Parts, NLTK                          | ID/Triplet/pair-wise Loss      | 59.00 | 79.49          | 86.60 |       |                |       |       |       |       |
| CA [89]          | Sensors20        | TI            | Global      | ResNet-50                | Bi-LSTM           | Cubic attention                      | ID/Ranking Loss                | 60.73 | 78.63          | 84.96 |       |                |       |       |       |       |
| CMKA [96]        | IEEE TIP21       | TI, IT        | Global      | ResNet50                 | Bi-LSTM           | /                                    | KL divergence, ID Loss         | 54.69 | 73.65          | 81.86 |       |                |       |       |       |       |
| ITMeetsAL [97]   | PR21             | TI, IT        | Global      | ResNet-152               | Bi-LSTM           | Shannon theory, adversarial learning | ID/Triplet Loss, KL divergence | 55.72 | 76.15          | 84.26 |       |                |       |       |       |       |
| DME [83]         | Neuro-comuting21 | TI            | Multi-scale | ResNet50                 | Bi-LSTM           | Attention, Parts                     | ID/CMM/RSP/CMR Loss            | 56.32 | 77.23          | 84.71 |       |                |       |       |       |       |
| CMMT [36]        | ICCV21           | TI            | Global      | ResNet-50                | Bi-LSTM           | Clustering                           | CMPM Loss                      | 57.10 | 78.14          | 85.23 |       |                |       |       |       |       |

| Method       | Publication      | Text-to-Image | Feature     | Visual backbone    | Language backbone | Auxiliary strategy                    | Alignment loss              | R@1   | CUHK-PEDES<br>R@5 | R@10  | R@1   | ICFG-PEDES<br>R@5 | R@10  | R@1   | RSTP Reid<br>R@5 | R@10  |
|--------------|------------------|---------------|-------------|--------------------|-------------------|---------------------------------------|-----------------------------|-------|-------------------|-------|-------|-------------------|-------|-------|------------------|-------|
| AMEN [106]   | PRCV21           | TI            | Global      | ResNet-50          | Bi-GRU            | Graph, GAN, Autoencoder               | Triplet/Adversarial Loss    | 57.16 | 78.64             | 86.22 |       |                   |       | 38.45 | 62.40            | 73.80 |
| T-MRS [77]   | IEEE TCST21      | TI            | Multi-scale | ResNet-101         | BERT              | Vanilla/Overlapped/Keypoints slicing  | ID/Contrastive Loss         | 57.67 | 78.25             | 84.93 |       |                   |       |       |                  |       |
| NAFS [82]    | arXiv21          | TI            | Multi-scale | ResNet-50          | BERT              | Parts, Contextual non-local attention | CMPM/CSAL Loss              | 59.94 | 79.86             | 86.70 |       |                   |       |       |                  |       |
| DSSL [31]    | ACM MM21         | TI            | Global      | ResNet-50          | Bi-GRU            | Attention, NLTK                       | ID/Alignment/MEC loss       | 59.98 | 80.41             | 87.56 |       |                   |       | 39.05 | 62.60            | 73.95 |
| MGEL [80]    | IJCAI21          | TI            | Multi-scale | ResNet-50          | Bi-LSTM           | Multi-head attention                  | ID/Triplet Loss             | 60.27 | 80.01             | 86.74 |       |                   |       |       |                  |       |
| SSAN [70]    | arXiv21          | TI            | Multi-scale | ResNet-50          | Bi-LSTM           | Regions, MASK                         | ID/CR Loss                  | 61.37 | 80.15             | 86.73 | 54.23 | 72.63             | 79.53 | 43.50 | 67.80            | 77.15 |
| TBPS [35]    | BMVC21           | TI, IT        | Global      | ResNet-101         | CLIP-TE/Bi-GRU    | CLIP                                  | ID/Contrastive Loss         | 64.08 | 81.73             | 88.19 |       |                   |       |       |                  |       |
| iVAD [104]   | Neuro-comuting22 | TI            | Multi-scale | VGG-16             | Bi-LSTM           | Virtual Attributes decoupling         | CMC/clc/ipe Loss            | 58.35 | 78.22             | 84.84 |       |                   |       |       |                  |       |
| SUM [115]    | KBS22            | TI            | global      | ResNet-50          | Bi-GRU            | Attention, NLTK                       | ID/Ranking Loss             | 59.22 | 80.35             | 87.60 |       |                   |       | 41.38 | 67.48            | 76.48 |
| CFLT [120]   | IEEE TIP22       | TI            | Multi-scale | ResNet-50          | BERT              | RVCF, Attention                       | CMPM Loss                   | 60.10 | 79.60             | 86.34 |       |                   |       |       |                  |       |
| LBUL [116]   | ACM MM22         | TI            | Global      | ResNet-50          | Bi-GRU            | /                                     | ID/Ranking Loss             | 61.95 | 81.16             | 87.19 |       |                   |       | 43.35 | 66.85            | 76.50 |
| ACSA [78]    | IEEE TMM22       | TI            | Multi-scale | Swin transformer   | BERT              | Parts                                 | ACSA/CMPM-C Loss            | 63.56 | 81.40             | 87.70 |       |                   |       | 48.40 | 71.85            | 81.45 |
| TIPCB [8]    | Neuro-comuting22 | TI            | Global      | ResNet-50          | BERT              | Mask                                  | CMPM Loss                   | 63.63 | 82.82             | 89.10 | 54.96 | 74.72             | 81.89 |       |                  |       |
| ISANet [81]  | arXiv22          | TI            | Multi-scale | ResNet-50          | Bi-LSTM           | Implicit local alignment              | ID/Ranking/Triplet Loss     | 63.92 | 82.15             | 87.69 | 57.73 | 75.42             | 81.72 |       |                  |       |
| SRCF [99]    | ECCV22           | TI            | Multi-scale | ResNet-50          | BERT              | Denosing filter                       | ID/CR/SEP Loss              | 64.04 | 82.99             | 88.81 | 57.18 | 75.01             | 81.49 |       |                  |       |
| SCFC [84]    | ArXiv22          | TI            | Multi-scale | ResNet-50          | BERT              | GCN, GAN                              | CMPM loss                   | 64.12 | 82.76             | 88.65 |       |                   |       | 45.88 | 70.45            | 81.30 |
| SAF [119]    | ICASSP22         | TI            | Global      | Vision Transformer | BERT              | Multi-head attention                  | CMPM Loss, KL divergence    | 64.13 | 82.62             | 88.40 |       |                   |       |       |                  |       |
| LGUR [118]   | ACM MM22         | TI            | Global      | DeiT-Small         | Bi-LSTM           | Mask, BERT                            | ID/Ranking Loss             | 64.21 | 81.94             | 87.93 |       |                   |       |       |                  |       |
| CAIBC [98]   | ACM MM22         | TI            | Global      | ResNet-50          | BERT              | Greyscale, Color mask                 | Triplet/Ranking/ID Loss     | 64.43 | 82.87             | 88.37 | 58.44 | 78.2              | 85.46 | 47.35 | 69.55            | 79.00 |
| AXM-Net [90] | AAAI22           | TI            | Multi-scale | ResNet-50          | BERT              | Parts, Self-Attention                 | ID/Triplet/Affinity Loss    | 64.44 | 80.52             | 86.77 | 49.37 | 71.69             | 79.27 | 53.14 | 74.03            | 81.95 |
| PMAN [79]    | PRCV22           | TI            | Multi-scale | ResNet-50          | BERT              | Multi-scale attention                 | CMPM Loss                   | 64.51 | 83.14             | 89.15 |       |                   |       |       |                  |       |
| PDA [2]      | IJCNN22          | TI            | Global      | ResNet-50          | BERT              | Mask, PCB                             | CMPM Loss                   | 65.26 | 84.58             | 89.98 |       |                   |       |       |                  |       |
| IVT [94]     | ECCVW22          | TI            | Global      | Vision Transformer | Transformer       | Mask, CLIP, NLTK                      | CMPM Loss                   | 65.59 | 83.11             | 89.21 | 56.04 | 73.60             | 80.22 | 46.70 | 70.00            | 78.80 |
| TFAF [85]    | IEEE SPL22       | TI            | Multi-scale | Pyramid ViT        | BERT, CNN         | Self Attention                        | CMPM/Reconstruction Loss    | 65.69 | 84.75             | 89.93 |       |                   |       |       |                  |       |
| C2A2 [34]    | ACM MM22         | TI            | Multi-scale | ResNet-50          | BERT              | Attribute dictionary                  | ID/CMPC Loss, KL divergence | 67.94 | 86.86             | 91.87 | 53.14 | 74.03             | 81.95 | 54.30 | 78.70            | 86.60 |

| Method         | Publication     | Text-to-Image | Feature     | Visual backbone    | Language backbone | Auxiliary strategy    | Alignment loss                  | R@1   | CUHK-PEDES<br>R@5 | R@10  | R@1   | ICFG-PEDES<br>R@5 | R@10  | R@1   | RSTP Reid<br>R@5 | R@10  |
|----------------|-----------------|---------------|-------------|--------------------|-------------------|-----------------------|---------------------------------|-------|-------------------|-------|-------|-------------------|-------|-------|------------------|-------|
| TGDA [114]     | IEEE TCSVT23    | TI            | Local       | ResNet-50          | BERT              | Mask, Attention, NLTK | ID/PKL/CM Loss                  | 64.64 | 83.38             | 89.34 | 57.26 | 75.19             | 81.80 | 48.35 | 73.15            | 80.30 |
| MSN-BRR [121]  | ACM MM23        | TI            | Multi-scale | ResNet-50          | BERT              | Attention             | RA/ER/ID Loss                   | 65.93 | 83.94             | 90.15 | 57.26 | 75.19             | 81.80 | 48.35 | 73.15            | 80.30 |
| CFine [38]     | IEEE TIP23      | TI, IT        | Multi-scale | CIIP-ViT           | BERT              | CLIP                  | CMPM-CMPC/Triplet Loss          | 69.67 | 85.93             | 91.15 | 60.83 | 76.55             | 82.42 | 50.55 | 72.50            | 81.60 |
| TP-TPS [39]    | arXiv23         | TI            | Local       | CLIP-ViT           | CLIP-Xformer      | CLIP                  | Ranking/Attribute Loss          | 70.16 | 86.10             | 90.98 | 54.12 | 75.45             | 82.97 | 54.12 | 75.45            | 82.97 |
| IRRA [40]      | CVPR23          | TI            | Multi-scale | CLIP-ViT           | CLIP-Xformer      | CLIP, Cross attention | ID/SDM/DRR Loss                 | 73.38 | 89.90             | 93.71 | 63.46 | 80.24             | 85.82 | 60.20 | 81.30            | 88.20 |
| RaSa [123]     | IJCAI23         | TI            | Global      | ALBEF-V            | ALBEF-T           | ALBEF                 | RA/SA/Contrastive Loss          | 76.51 | 90.29             | 94.25 | 65.28 | 80.40             | 85.12 | 66.90 | 86.50            | 91.25 |
| APTМ [37]      | ACM MM23        | TI            | Global      | Swin Transformer   | BERT              | Attributes, Mask      | ID/Contrastive Loss             | 76.53 | 90.04             | 94.15 | 68.51 | 82.99             | 87.56 | 67.50 | 85.70            | 91.45 |
| PFM-EKFP [122] | arXiv23         | TI            | Global      | Vision Transformer | BERT              | Cross attention       | CMPM/CMPC Loss                  | 77.24 | 93.71             | 96.98 | 69.29 | 89.10             | 94.06 | 48.65 | 77.15            | 87.00 |
| CCL [41]       | ESWA24          | TI            | Global      | CLIP-ViT           | CLIP-Xformer      | CLIP                  | CMPM Loss                       | 67.25 | 86.10             | 91.45 | 58.33 | 76.75             | 83.38 | 51.30 | 75.25            | 84.60 |
| CANC [42]      | Inform Fusion24 | TI            | Global      | CLIP-ViT           | CLIP-Xformer      | Clustering            | KL divergence, Contrastive Loss | 69.61 | 88.10             | 91.47 | 60.52 | 78.36             | 84.13 | 56.24 | 80.15            | 86.61 |
| RDE [43]       | CVPR24          | TI            | Global      | CLIP-ViT           | CLIP-Xformer      | Mask                  | Triplet Alignment Loss          | 71.33 | 87.41             | 91.81 | 63.76 | 79.53             | 84.91 | 62.85 | 83.20            | 89.15 |
| CFAM [44]      | CVPR24          | TI            | Global      | CLIP-ViT           | CLIP-Xformer      | CLIP                  | ID/GA/LA Loss                   | 75.60 | 90.53             | 94.36 | 65.38 | 81.17             | 86.35 | 62.45 | 83.55            | 91.10 |
| AUL [45]       | AAAI24          | TI            | Global      | Swin Transformer   | BERT              | Mask                  | CMM Loss                        | 77.23 | 90.43             | 94.41 | 69.16 | 83.32             | 88.37 | 71.65 | 87.55            | 92.05 |

## References

- [1] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, S. C. Hoi, Deep learning for person re-identification: A survey and outlook, *IEEE transactions on pattern analysis and machine intelligence* 44 (2021) 2872–2893.
- [2] H.-Q. Cai, X. Li, Y. Ji, Y. Li, C.-P. Liu, Parallel data augmentation for text-based person re-identification, in: *2022 International Joint Conference on Neural Networks (IJCNN)*, IEEE, 2022, pp. 1–8.
- [3] Z. Wang, Z. Wang, Y. Zheng, Y. Wu, W. Zeng, S. Satoh, Beyond intra-modality: A survey of heterogeneous person re-identification, *arXiv preprint arXiv:1905.10048* (2019).
- [4] S. Xiang, Y. Fu, G. You, T. Liu, Unsupervised Domain Adaptation Through Synthesis For Person Re-Identification, in: *2020 IEEE International Conference on Multimedia and Expo (ICME)*, 2020, pp. 1–6. doi:10.1109/ICME46284.2020.9102822.
- [5] S. Xiang, D. Qian, M. Guan, B. Yan, T. Liu, Y. Fu, G. You, Less Is More: Learning from Synthetic Data with Fine-Grained Attributes for Person Re-Identification, *ACM Transactions on Multimedia Computing, Communications, and Applications* 19 (2023) 173:1–173:20.
- [6] X. Qian, Y. Fu, T. Xiang, W. Wang, J. Qiu, Y. Wu, Y.-G. Jiang, X. Xue, Pose-Normalized Image Generation for Person Re-identification, in: V. Ferrari, M. Hebert, C. Sminchisescu, Y. Weiss (Eds.), *Computer Vision – ECCV 2018*, Springer International Publishing, Cham, 2018, pp. 661–678. doi:10.1007/978-3-030-01240-3\_40.
- [7] S. Li, T. Xiao, H. Li, B. Zhou, D. Yue, X. Wang, Person search with natural language description, in: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1970–1979.
- [8] Y. Chen, G. Zhang, Y. Lu, Z. Wang, Y. Zheng, Tipcb: A simple but effective part-based convolutional baseline for text-based person search, *Neurocomputing* 494 (2022) 171–181.
- [9] M. Hirzer, C. Beleznaï, P. M. Roth, H. Bischof, Person re-identification by descriptive and discriminative classification, in: *Image Analysis: 17th Scandinavian Conference, SCIA 2011, Ystad, Sweden, May 2011. Proceedings 17*, Springer, 2011, pp. 91–102.
- [10] F. Xiong, M. Gou, O. Camps, M. Szaier, Person re-identification using kernel-based metric learning methods, in: *ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part VII 13*, Springer, 2014, pp. 1–16.
- [11] Y. Lin, L. Zheng, Z. Zheng, Y. Wu, Z. Hu, C. Yan, Y. Yang, Improving person re-identification by attribute and identity learning, *Pattern recognition* 95 (2019) 151–161.
- [12] Y. Deng, P. Luo, C. C. Loy, X. Tang, Pedestrian attribute recognition at far distance, in: *Proceedings of the 22nd ACM international conference on Multimedia*, 2014, pp. 789–792.
- [13] R. Layne, T. M. Hospedales, S. Gong, Q. Mary, Person re-identification by attributes, in: *British Machine Vision Conference (BMVC)*, volume 2, 2012, p. 8.
- [14] X. Wang, S. Zheng, R. Yang, A. Zheng, Z. Chen, J. Tang, B. Luo, Pedestrian attribute recognition: A survey, *Pattern Recognition* 121 (2022) 108220.
- [15] Y.-T. Cao, J. Wang, D. Tao, Symbiotic adversarial learning for attribute-based person search, in: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, Springer, 2020, pp. 230–247.
- [16] Z. Yin, W.-S. Zheng, A. Wu, H.-X. Yu, H. Wan, X. Guo, F. Huang, J. Lai, Adversarial attribute-image person re-identification, in: *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, AAAI Press, 2018, pp. 1100–1106.
- [17] Y. Jing, C. Si, J. Wang, W. Wang, L. Wang, T. Tan, Pose-guided multi-granularity attention network for text-based person search, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 2020, pp. 11189–11196.
- [18] Q. Dong, S. Gong, X. Zhu, Person search by text attribute query as zero-shot learning, in: *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 3652–3661.
- [19] B. Jeong, J. Park, S. Kwak, Asmr: Learning attribute-based person search with adaptive semantic margin regularizer, in: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 12016–12025.
- [20] S. Li, T. Xiao, H. Li, W. Yang, X. Wang, Identity-aware textual-visual matching with latent co-attention, in: *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 1890–1899.
- [21] K. Zheng, W. Liu, J. Liu, Z.-J. Zha, T. Mei, Hierarchical gumbel attention network for text-based person search, in: *Proceedings of the 28th ACM international conference on multimedia*, 2020, pp. 3441–3449.
- [22] Z. Zheng, L. Zheng, M. Garrett, Y. Yang, M. Xu, Y.-D. Shen, Dual-path convolutional image-text embeddings with instance loss, *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 16 (2020) 1–23.
- [23] Y. Jing, C. Si, J. Wang, W. Wang, L. Wang, T. Tan, Pose-guided joint global and attentive local matching network for text-based person search, *Association for the Advance of Artificial Intelligence (AAAI)* (2020).
- [24] Z. Wang, Z. Fang, J. Wang, Y. Yang, Vitaa: Visual-textual attributes alignment in person search by natural language, in: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*, Springer, 2020, pp. 402–420.
- [25] K. Niu, Y. Huang, W. Ouyang, L. Wang, Improving description-based person re-identification by multi-granularity image-text alignments, *IEEE Transactions on Image Processing* 29 (2020) 5542–5556.
- [26] D. Chen, H. Li, X. Liu, Y. Shen, J. Shao, Z. Yuan, X. Wang, Improving deep visual representation for person re-identification by global and local image-language association, in: *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 54–70.
- [27] Y. Zhang, H. Lu, Deep cross-modal projection learning for image-text matching, in: *Proceedings of the European conference on computer vision (ECCV)*, Springer International Publishing, 2018, pp. 686–701.
- [28] T. Chen, C. Xu, J. Luo, Improving text-based person search by spatial matching and adaptive threshold, in: *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, IEEE, 2018, pp. 1879–1887.
- [29] N. Sarafianos, X. Xu, I. A. Kakadiaris, Adversarial representation learning for text-to-image matching, in: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 5814–5824.
- [30] J. Liu, Z.-J. Zha, R. Hong, M. Wang, Y. Zhang, Deep adversarial graph attention convolution network for text-based person search, in: *Proceedings of the 27th ACM International Conference on Multimedia*, Association for Computing Machinery, New York, NY, USA, 2019, pp. 665–673.
- [31] A. Zhu, Z. Wang, Y. Li, X. Wan, J. Jin, T. Wang, F. Hu, G. Hua, Dssl: Deep surroundings-person separation learning for text-based person retrieval, in: *Proceedings of the 29th ACM international conference on multimedia*, Association for Computing Machinery, 2021, pp. 209–217.
- [32] Y. Wu, Z. Yan, X. Han, G. Li, C. Zou, S. Cui, Lapscore: language-guided person search via color reasoning, in: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1624–1633.
- [33] S. Yan, Y. Zhang, M. Xie, D. Zhang, Z. Yu, Cross-domain person re-identification with pose-invariant feature decomposition and hypergraph structure alignment, *Neurocomputing* 467 (2022) 229–241.
- [34] K. Niu, L. Huang, Y. Huang, P. Wang, L. Wang, Y. Zhang, Cross-modal co-occurrence attributes alignments for person search by language, in: *Proceedings of the 30th ACM international conference on multimedia*, 2022, pp. 4426–4434.
- [35] X. Han, S. He, L. Zhang, T. Xiang, Text-based person search with limited data, *arXiv preprint arXiv:2110.10807* (2021).
- [36] S. Zhao, C. Gao, Y. Shao, W.-S. Zheng, N. Sang, Weakly supervised text-based person re-identification, in: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp.

- 11395–11404.
- [37] S. Yang, Y. Zhou, Z. Zheng, Y. Wang, L. Zhu, Y. Wu, Towards unified text-based person retrieval: A large-scale multi-attribute and language search benchmark, in: Proceedings of the 31st ACM International Conference on Multimedia, 2023, pp. 4492–4501.
- [38] S. Yan, N. Dong, L. Zhang, J. Tang, Clip-driven fine-grained text-image person re-identification, *IEEE Transactions on Image Processing* (2023).
- [39] G. Wang, F. Yu, J. Li, Q. Jia, S. Ding, Exploiting the textual potential from vision-language pre-training for text-based person search, *arXiv preprint arXiv:2303.04497* (2023).
- [40] D. Jiang, M. Ye, Cross-modal implicit relation reasoning and aligning for text-to-image person retrieval, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 2787–2797.
- [41] G. Du, T. Gong, L. Zhang, Contrastive completing learning for practical text-image person ReID: Robuster and cheaper, *Expert Systems with Applications* 248 (2024) 123399.
- [42] T. Gong, J. Wang, L. Zhang, Cross-modal semantic aligning and neighbor-aware completing for robust text-image person retrieval, *Information Fusion* 112 (2024) 102544.
- [43] Y. Qin, Y. Chen, D. Peng, X. Peng, J. T. Zhou, P. Hu, Noisy-Correspondence Learning for Text-to-Image Person Re-Identification, in: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE Computer Society, 2024, pp. 27187–27196. doi:10.1109/CVPR52733.2024.02568.
- [44] J. Zuo, H. Zhou, Y. Nie, F. Zhang, T. Guo, N. Sang, Y. Wang, C. Gao, UFineBench: Towards Text-based Person Retrieval with Ultra-fine Granularity, in: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE Computer Society, 2024, pp. 22010–22019. doi:10.1109/CVPR52733.2024.02078.
- [45] S. Li, C. He, X. Xu, F. Shen, Y. Yang, H. T. Shen, Adaptive Uncertainty-Based Learning for Text-Based Person Retrieval, Proceedings of the AAAI Conference on Artificial Intelligence 38 (2024) 3172–3180.
- [46] A. Bedagkar-Gala, S. K. Shah, A survey of approaches and trends in person re-identification, *Image and vision computing* 32 (2014) 270–286.
- [47] L. Zheng, L. Shen, L. Tian, S. Wang, J. Bu, Q. Tian, Person re-identification meets image search, *arXiv preprint arXiv:1502.02171* (2015).
- [48] H. Chahar, N. Nain, A study on deep convolutional neural network based approaches for person re-identification, in: Pattern Recognition and Machine Intelligence: 7th International Conference, PReMI 2017, Kolkata, India, December 5-8, 2017, Proceedings 7, Springer, 2017, pp. 543–548.
- [49] K. Wang, H. Wang, M. Liu, X. Xing, T. Han, Survey on person re-identification based on deep learning, *CAAI Transactions on Intelligence Technology* 3 (2018) 219–227.
- [50] D. Wu, H. Huang, Q. Zhao, S. Zhang, J. Qi, J. Hu, Overview of deep learning based pedestrian attribute recognition and re-identification, *Heliyon* 8 (2022).
- [51] N. Mathur, S. Mathur, D. Mathur, P. Dadheech, A brief survey of deep learning techniques for person re-identification, in: 2020 3rd International Conference on Emerging Technologies in Computer Engineering: Machine Learning and Internet of Things (ICETCE), IEEE, 2020, pp. 129–138.
- [52] K. Islam, Person search: New paradigm of person re-identification: A survey and outlook of recent works, *Image and Vision Computing* 101 (2020) 103970.
- [53] B. Lavi, I. Ullah, M. Fatan, A. Rocha, Survey on reliable deep learning-based person re-identification models: Are we there yet?, *arXiv preprint arXiv:2005.00355* (2020).
- [54] G. Zou, G. Fu, X. Peng, Y. Liu, M. Gao, Z. Liu, Person re-identification based on metric learning: a survey, *multimedia tools and applications* 80 (2021) 26855–26888.
- [55] C. Yang, F. Qi, H. Jia, Survey on unsupervised techniques for person re-identification, in: 2021 2nd International Conference on Computing and Data Science (CDS), IEEE, 2021, pp. 161–164.
- [56] E. Yaghoubi, A. Kumar, H. Proença, Sss-pr: A short survey of surveys in person re-identification, *Pattern Recognition Letters* 143 (2021) 50–57.
- [57] Y. Wang, S. Yang, S. Liu, Z. Zhang, Cross-domain person re-identification: a review, in: Artificial Intelligence in China: Proceedings of the 2nd International Conference on Artificial Intelligence in China, Springer, 2021, pp. 153–160.
- [58] Z. Ming, M. Zhu, X. Wang, J. Zhu, J. Cheng, C. Gao, Y. Yang, X. Wei, Deep learning-based person re-identification methods: A survey and outlook of recent works, *Image and Vision Computing* 119 (2022) 104394.
- [59] Y. Peng, J. Wu, B. Xu, C. Cao, X. Liu, Z. Sun, Z. He, Deep learning based occluded person re-identification: A survey, *ACM Transactions on Multimedia Computing, Communications and Applications* 20 (2023) 1–27.
- [60] O. Vinyals, A. Toshev, S. Bengio, D. Erhan, Show and tell: A neural image caption generator, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3156–3164.
- [61] A. Zahra, N. Perwaiz, M. Shahzad, M. M. Fraz, Person re-identification: A retrospective on domain specific open challenges and future trends, *Pattern Recognition* 142 (2023) 109669.
- [62] N. K. Singh, M. Khare, H. B. Jethva, A comprehensive survey on person re-identification approaches: various aspects, *Multimedia Tools and Applications* 81 (2022) 15747–15791.
- [63] N. Huang, J. Liu, Y. Miao, Q. Zhang, J. Han, Deep learning for visible-infrared cross-modality person re-identification: A comprehensive review, *Information Fusion* 91 (2023) 396–411.
- [64] S. Karanam, M. Gou, Z. Wu, A. Rates-Borras, O. Camps, R. J. Radke, A systematic evaluation and benchmark for person re-identification: Features, metrics, and datasets, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41 (2019) 523–536.
- [65] Y. Deng, P. Luo, C. C. Loy, X. Tang, Pedestrian attribute recognition at far distance, in: Proceedings of the 22nd ACM international conference on Multimedia, Association for Computing Machinery, 2014, pp. 789–792.
- [66] D. Li, Z. Zhang, X. Chen, H. Ling, K. Huang, A richly annotated dataset for pedestrian attribute recognition, *arXiv preprint arXiv:1603.07054* (2016).
- [67] D. Li, Z. Zhang, X. Chen, K. Huang, A richly annotated pedestrian dataset for person retrieval in real surveillance scenarios, *IEEE transactions on image processing* 28 (2018) 1575–1590.
- [68] X. Liu, H. Zhao, M. Tian, L. Sheng, J. Shao, S. Yi, J. Yan, X. Wang, Hydraplus-net: Attentive deep features for pedestrian analysis, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 350–359.
- [69] T. Li, J. Liu, W. Zhang, Y. Ni, W. Wang, Z. Li, Uav-human: A large benchmark for human behavior understanding with unmanned aerial vehicles, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 16266–16275.
- [70] Z. Ding, C. Ding, Z. Shao, D. Tao, Semantically self-aligned network for text-to-image part-aware person re-identification, *arXiv preprint arXiv:2107.12666* (2021).
- [71] D. A. Vaquero, R. S. Feris, D. Tran, L. Brown, A. Hampapur, M. Turk, Attribute-based people search in surveillance environments, in: 2009 workshop on applications of computer vision (WACV), IEEE, 2009, pp. 1–8.
- [72] W. Li, R. Zhao, T. Xiao, X. Wang, Deepreid: Deep filter pairing neural network for person re-identification, in: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 152–159.
- [73] T. Xiao, S. Li, B. Wang, L. Lin, X. Wang, Joint detection and identification feature learning for person search, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 3415–3424.
- [74] D. Gray, S. Brennan, H. Tao, Evaluating appearance models for recognition, reacquisition, and tracking, in: Proc. IEEE international workshop on performance evaluation for tracking and surveillance

- (PETS), volume 3, 2007, pp. 1–7.
- [75] W. Li, R. Zhao, X. Wang, Human reidentification with transferred metric learning, in: Computer Vision–ACCV 2012: 11th Asian Conference on Computer Vision, Daejeon, Korea, November 5-9, 2012, Revised Selected Papers, Part I 11, Springer, 2013, pp. 31–44.
- [76] L. Wei, S. Zhang, W. Gao, Q. Tian, Person transfer gan to bridge domain gap for person re-identification, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 79–88.
- [77] H. Li, J. Xiao, M. Sun, E. G. Lim, Y. Zhao, Transformer-based language-person search with multiple region slicing, IEEE Transactions on Circuits and Systems for Video Technology 32 (2021) 1624–1633.
- [78] Z. Ji, J. Hu, D. Liu, L. Y. Wu, Y. Zhao, Asymmetric cross-scale alignment for text-based person search, IEEE Transactions on Multimedia 25 (2022) 7699–7709.
- [79] Y. Wang, D. Qi, C. Zhao, Part-based multi-scale attention network for text-based person search, in: PRCV 2022, Springer, 2022, pp. 462–474.
- [80] C. Wang, Z. Luo, Y. Lin, S. Li, Text-based person search via multi-granularity embedding learning, in: Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, International Joint Conferences on Artificial Intelligence Organization, 2021, pp. 1068–1074.
- [81] S. Yan, H. Tang, L. Zhang, J. Tang, Image-specific information suppression and implicit local alignment for text-based person search, IEEE transactions on neural networks and learning systems (2023).
- [82] C. Gao, G. Cai, X. Jiang, F. Zheng, J. Zhang, Y. Gong, P. Peng, X. Guo, X. Sun, Contextual non-local alignment over full-scale representation for text-based person search, arXiv preprint arXiv:2101.03036 (2021).
- [83] C. Wang, Z. Luo, Z. Zhong, S. Li, Divide-and-merge the embedding space for cross-modality person search, Neurocomputing 463 (2021) 388–399.
- [84] F. Li, H. Zhou, H. Li, Y. Zhang, Z. Yu, Person text-image matching via text-feature interpretability embedding and external attack node implantation, IEEE Transactions on Emerging Topics in Computational Intelligence (2024).
- [85] S. Li, A. Lu, Y. Huang, C. Li, L. Wang, Joint token and feature alignment framework for text-based person search, IEEE Signal Processing Letters 29 (2022) 2238–2242.
- [86] Z. Wang, A. Zhu, Z. Zheng, J. Jin, Z. Xue, G. Hua, Img-net: inner-cross-modal attentional multigranular network for description-based person re-identification, Journal of Electronic Imaging 29 (2020) 043028–043028.
- [87] T.-Y. Liu, C. Zhu, L. Yang, Efficient text-based person search via single-stage identity-guided attribute parsing and alignment, in: 2022 26th International Conference on Pattern Recognition (ICPR), 2022, pp. 4111–4117.
- [88] Y. Jing, C. Si, J. Wang, W. Wang, L. Wang, T. Tan, Cascade attention network for person search: Both image and text-image similarity selection, arXiv preprint arXiv:1809.08440 2 (2018) 5.
- [89] Y. Li, H. Xu, J. Xiao, Hybrid attention network for language-based person search, Sensors 20 (2020) 5279.
- [90] A. Farooq, M. Awais, J. Kittler, S. S. Khalid, Axm-net: Implicit cross-modal feature alignment for person re-identification, in: Proceedings of the AAAI conference on artificial intelligence, volume 36, 2022, pp. 4477–4485.
- [91] K.-H. Lee, X. Chen, G. Hua, H. Hu, X. He, Stacked cross attention for image-text matching, in: Proceedings of the European conference on computer vision (ECCV), 2018, pp. 201–216.
- [92] S. Zhang, D. Cheng, W. Luo, Y. Xing, D. Long, H. Li, K. Niu, G. Liang, Y. Zhang, Text-based person search in full images via semantic-driven proposal generation, in: Proceedings of the 4th International Workshop on Human-centric Multimedia Analysis, 2023, pp. 5–14.
- [93] D. Chen, M. Wang, H. Chen, L. Wu, J. Qin, W. Peng, Cross-modal retrieval with heterogeneous graph embedding, in: Proceedings of the 30th ACM International Conference on Multimedia, Association for Computing Machinery, New York, NY, USA, 2022, pp. 3291–3300.
- [94] X. Shu, W. Wen, H. Wu, K. Chen, Y. Song, R. Qiao, B. Ren, X. Wang, See finer, see more: Implicit modality alignment for text-based person retrieval, in: European Conference on Computer Vision, Springer, 2022, pp. 624–641.
- [95] S. Reed, Z. Akata, H. Lee, B. Schiele, Learning deep representations of fine-grained visual descriptions, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 49–58.
- [96] Y. Chen, R. Huang, H. Chang, C. Tan, T. Xue, B. Ma, Cross-modal knowledge adaptation for language-based person search, IEEE Transactions on Image Processing 30 (2021) 4057–4069.
- [97] W. Chen, Y. Liu, E. M. Bakker, M. S. Lew, Integrating information theory and adversarial learning for cross-modal retrieval, Pattern Recognition 117 (2021) 107983.
- [98] Z. Wang, A. Zhu, J. Xue, X. Wan, C. Liu, T. Wang, Y. Li, Caibc: Capturing all-round information beyond color for text-based person retrieval, in: Proceedings of the 30th ACM international conference on multimedia, 2022, pp. 5314–5322.
- [99] W. Suo, M. Sun, K. Niu, Y. Gao, P. Wang, Y. Zhang, Q. Wu, A simple and robust correlation filtering method for text-based person search, in: European conference on computer vision, Springer, 2022, pp. 726–742.
- [100] S. Aggarwal, V. B. Radhakrishnan, A. Chakraborty, Text-based person search via attribute-aided matching, in: Proceedings of the IEEE/CVF winter conference on applications of computer vision, 2020, pp. 2617–2625.
- [101] A. Farooq, M. Awais, F. Yan, J. Kittler, A. Akbari, S. S. Khalid, A convolutional baseline for person re-identification using vision and language descriptions, arXiv preprint arXiv:2003.00808 (2020).
- [102] A. Farooq, M. Awais, J. Kittler, A. Akbari, S. S. Khalid, Cross modal person re-identification with visual-textual queries, in: 2020 IEEE International Joint Conference on Biometrics (IJCB), IEEE, 2020, pp. 1–8.
- [103] J. Ge, G. Gao, Z. Liu, Visual-textual association with hardest and semi-hard negative pairs mining for person search, arXiv preprint arXiv:1912.03083 (2019).
- [104] C. Wang, Z. Luo, Y. Lin, S. Li, Improving embedding learning by virtual attribute decoupling for text-based person search, Neural Computing and Applications (2022) 1–23.
- [105] Y. Wang, C. Bo, D. Wang, S. Wang, Y. Qi, H. Lu, Language person search with mutually connected classification loss, in: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2019, pp. 2057–2061.
- [106] Z. Wang, J. Xue, A. Zhu, Y. Li, M. Zhang, C. Zhong, Amen: Adversarial multi-space embedding network for text-based person re-identification, in: PRCV 2021, Springer, 2021, pp. 462–473.
- [107] K. Niu, Y. Huang, L. Wang, Fusing two directions in cross-domain adaptation for real life person search by language, in: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, 2019, pp. 0–0.
- [108] K. Niu, Y. Huang, L. Wang, Textual dependency embedding for person search by language, in: Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 4032–4040.
- [109] T. Ma, M. Yang, H. Rong, Y. Qian, Y. Tian, N. Al-Nabhan, Dual-path cnn with max gated block for text-based person re-identification, Image and Vision Computing 111 (2021) 104168.
- [110] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556 (2014).
- [111] A. G. Howard, Mobilenets: Efficient convolutional neural networks for mobile vision applications, arXiv preprint arXiv:1704.04861 (2017).
- [112] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

- [113] Y. Jing, W. Wang, L. Wang, T. Tan, Cross-modal cross-domain moment alignment network for person search, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 10678–10686.
- [114] L. Gao, K. Niu, B. Jiao, P. Wang, Y. Zhang, Addressing information inequality for text-based person search via pedestrian-centric visual denoising and bias-aware alignments, *IEEE Transactions on Circuits and Systems for Video Technology* 33 (2023) 7884–7899.
- [115] Z. Wang, A. Zhu, J. Xue, D. Jiang, C. Liu, Y. Li, F. Hu, Sum: Serialized updating and matching for text-based person retrieval, *Knowledge-Based Systems* 248 (2022) 108891.
- [116] Z. Wang, A. Zhu, J. Xue, X. Wan, C. Liu, T. Wang, Y. 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, 2022, pp. 1984–1992.
- [117] J. D. M.-W. C. Kenton, L. K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of naacL-HLT, volume 1, Minneapolis, Minnesota, 2019, p. 2.
- [118] Z. Shao, X. Zhang, M. Fang, Z. Lin, J. Wang, C. Ding, Learning granularity-unified representations for text-to-image person re-identification, in: Proceedings of the 30th acm international conference on multimedia, 2022, pp. 5566–5574.
- [119] S. Li, M. Cao, M. Zhang, Learning semantic-aligned feature representation for text-based person search, in: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2022, pp. 2724–2728.
- [120] C. Gao, G. Cai, X. Jiang, F. Zheng, J. Zhang, Y. Gong, F. Lin, X. Sun, X. Bai, Conditional feature learning based transformer for text-based person search, *IEEE Transactions on Image Processing* 31 (2022) 6097–6108.
- [121] S. Li, X. Xu, F. Shen, Y. Yang, Multi-granularity separation network for text-based person retrieval with bidirectional refinement regularization, in: Proceedings of the 2023 ACM International Conference on Multimedia Retrieval, 2023, pp. 307–315.
- [122] H. Li, S. Yang, Y. Zhang, D. Tao, Z. Yu, Progressive feature mining and external knowledge-assisted text-pedestrian image retrieval, *arXiv preprint arXiv:2308.11994* (2023).
- [123] Y. Bai, M. Cao, D. Gao, Z. Cao, C. Chen, Z. Fan, L. Nie, M. Zhang, Rasa: Relation and sensitivity aware representation learning for text-based person search, *arXiv preprint arXiv:2305.13653* (2023).
- [124] J. Li, R. Selvaraju, A. Gotmare, S. Joty, C. Xiong, S. C. H. Hoi, Align before fuse: Vision and language representation learning with momentum distillation, *Advances in neural information processing systems* 34 (2021) 9694–9705.
- [125] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, in: International Conference on Learning Representations, 2023.
- [126] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International conference on machine learning, PMLR, 2021, pp. 8748–8763.
- [127] Z. Zheng, L. Zheng, Y. Yang, A discriminatively learned cnn embedding for person reidentification, *ACM transactions on multimedia computing, communications, and applications (TOMM)* 14 (2017) 1–20.
- [128] Y. Duan, J. Lu, J. Feng, J. Zhou, Deep localized metric learning, *IEEE Transactions on Circuits and Systems for Video Technology* 28 (2017) 2644–2656.
- [129] L. Zheng, Y. Yang, A. G. Hauptmann, Person re-identification: Past, present and future, *arXiv preprint arXiv:1610.02984* (2016).
- [130] C. Luo, Y. Chen, N. Wang, Z. Zhang, Spectral feature transformation for person re-identification, in: Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 4976–4985.
- [131] M. Zheng, S. Karanam, Z. Wu, R. J. Radke, Re-identification with consistent attentive siamese networks, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 5735–5744.
- [132] S. Chopra, R. Hadsell, Y. LeCun, Learning a similarity metric discriminatively, with application to face verification, in: 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05), volume 1, IEEE, 2005, pp. 539–546.
- [133] Y. Wang, Z. Chen, F. Wu, G. Wang, Person re-identification with cascaded pairwise convolutions, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 1470–1478.
- [134] R. R. Varior, B. Shuai, J. Lu, D. Xu, G. Wang, A siamese long short-term memory architecture for human re-identification, in: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII 14, Springer, 2016, pp. 135–153.
- [135] G. Chechik, V. Sharma, U. Shalit, S. Bengio, Large scale online learning of image similarity through ranking., *Journal of Machine Learning Research* 11 (2010).
- [136] A. Hermans, L. Beyer, B. Leibe, In defense of the triplet loss for person re-identification, *arXiv preprint arXiv:1703.07737* (2017).
- [137] S. Iodice, K. Mikolajczyk, Text attribute aggregation and visual feature decomposition for person search., in: British Machine Vision Conference (BMVC), 2020.
- [138] W. Chen, X. Chen, J. Zhang, K. Huang, Beyond triplet loss: a deep quadruplet network for person re-identification, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 403–412.
- [139] S. Zhang, Y. Wang, T. Chai, A. Li, A. K. Jain, Realgait: Gait recognition for person re-identification, *arXiv preprint arXiv:2201.04806* (2022).
- [140] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International conference on machine learning, PMLR, 2021, pp. 8748–8763.
- [141] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer, High-resolution image synthesis with latent diffusion models, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 10684–10695.
- [142] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5–9, 2015, proceedings, part III 18, Springer, 2015, pp. 234–241.
- [143] L. Zhang, A. Rao, M. Agrawala, Adding conditional control to text-to-image diffusion models, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 3836–3847.
- [144] Z. Cao, T. Simon, S.-E. Wei, Y. Sheikh, Realtime multi-person 2d pose estimation using part affinity fields, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 7291–7299.
- [145] T. B. Brown, Language models are few-shot learners, *arXiv preprint arXiv:2005.14165* (2020).
- [146] A. Hertz, R. Mokady, J. Tenenbaum, K. Aberman, Y. Pritch, D. Cohen-Or, Prompt-to-prompt image editing with cross attention control, *arXiv preprint arXiv:2208.01626* (2022).
- [147] Y. Sun, L. Zheng, Y. Yang, Q. Tian, S. Wang, Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline), in: Proceedings of the European conference on computer vision (ECCV), 2018, pp. 480–496.
- [148] X. Pan, P. Luo, J. Shi, X. Tang, Two at once: Enhancing learning and generalization capacities via ibn-net, in: Proceedings of the european conference on computer vision (ECCV), 2018, pp. 464–479.