Contrastive Completing Learning For Practical Text-Image person ReID: Robuster and Cheaper

Guodong Du, Tiantian Gong, Liyan Zhang

***Abstract*—Text-image person re-identification (TIReID) seeks to leverage textual descriptions for the retrieval of target pedestrians. Due to its versatility, TIReID has gained increasing attention. However, manual annotation of textual descriptions and identity labels can be time-consuming and costly, limiting its scalability in practical settings. Privacy concerns and poor data storage can lead to data loss or ineffectiveness, further exacerbating challenges in real-world scenarios. To address these limitations, we propose for the first time incomplete Text-image person re-identification (iTIReID), which comprises a small amount of complete pairwise data and a large amount of incomplete data, where all identity labels are unavailable. We introduce a novel Contrastive Completing Learning (CCL) framework for iTIReID, consisting of two stages: Pure Contrastive Learning (PCL) and Feature Completion Contrastive Learning (FCCL). In PCL, only complete pairwise data is utilized for training, which serves as a preliminary improvement of the model's capacity and prepares for the upcoming feature completion stage. In FCCL, available features are used to complete missing modality features and facilitate effective training with incomplete data. During this process, Cross-modal Semantic Measure (CSM) is proposed to leverage intra-modality similarity to measure cross-modal similarity and filter out features with the highest semantic similarity, thereby circumventing modality discrepancy. Semantic-Weighted Generation (SWG) is proposed to generate approximate features based on the semantic similarity weight of the similar features. To fully leverage pairwise data for label-free training, we introduce the contrastive CMPM (CCMPM) loss for contrastive learning to achieve weakly supervised training. Experimental results verify the effectiveness of our proposed methods and demonstrate competitive performance compared to fully supervised methods using complete data.**

***Index Terms*—Person re-identification, image-text retrieval, weakly supervised learning, modality imbalance, robustness.**

I. Introduction

PERSON re-identification (ReID) is a highly challenging task in the field of computer vision that has garnered significant research attention. Over the past decade, ReID technologies have made great strides in practical applications, but they rely on the assumption that target pedestrian images can be captured by different disjoint cameras [1], [2]. This assumption may not hold true in actual scenarios due to missing road monitoring or target pedestrian occlusion. As a result, researchers have developed alternative approaches such as text-image person ReID (TIReID), which relies on verbal descriptions provided by witnesses at the scene to search for the target pedestrian [3], [4]. In practical applications, language descriptions are often easier to acquire and more flexible to utilize than images, making TIReID research extremely valuable. Currently, TIReID is receiving increasing attention both from academia and industry.

In comparison to typical text-image retrieval tasks, TIReID is especially challenging due to its dual focus on person re-identification and cross-modal retrieval. As a result, it is considered a fine-grained cross-modal retrieval task that requires addressing both the modality discrepancy between queries and retrieval candidates and the need to explore and learn distinctive features to achieve optimal results. To achieve these goals, many existing methods rely on multi-scale learning or cross-modal attention mechanisms to map textual and visual modalities into a shared feature space while also ensuring semantic alignment across different scales. For feature processing, these methods often rely on breaking down textual descriptions and images to obtain fine-grained features [5]–[7]. In metric learning, the majority of outstanding methods rely on identity loss to reduce intra-class variance while further optimizing the feature space using ranking loss [7], [8]. Thanks to the precise identity labels in existing datasets, detailed image descriptions, and excellent training methods, these methods have achieved promising results.

Although TIReID has made significant progress, there are still two critical issues that limit its scalability in real-world scenarios. The first issue concerns the high cost of data acquisition. Training TIReID models requires a large number of images, detailed descriptions, and annotated identity labels. Annotating identity labels in cross-camera scenarios is notoriously time-consuming and labor-intensive. Moreover, to ensure accurate and objective descriptions, multiple annotators are often required for annotation or correction, increasing the complexity and workload compared to obtaining pedestrian labels. Compared to image-based ReID, obtaining a complete and correct TIReID large-scale dataset typically incurs much higher time and financial costs, which makes it challenging to deploy in real-world applications. As a result, using a small number of samples with complete descriptions and a larger number of raw samples without descriptions to train and obtain robust models has become an attractive and promising research direction, although relevant studies have not yet appeared.

The second key issue that hinders the application of TIReID in real-world scenarios is the robustness of the model. Previous TIReID studies have assumed that different modality data are balanced, meaning that one modality has corresponding samples for each sample in another modality. However, in real-world scenarios, privacy issues may make data difficult to acquire, and poor storage can result in data loss or invalidity. These factors can lead to missing data. As a result, not all the data in a collected dataset may be ideally paired, which can cause modality imbalance problems. Due to insufficient and incomplete data, semantic information cannot be well aligned,

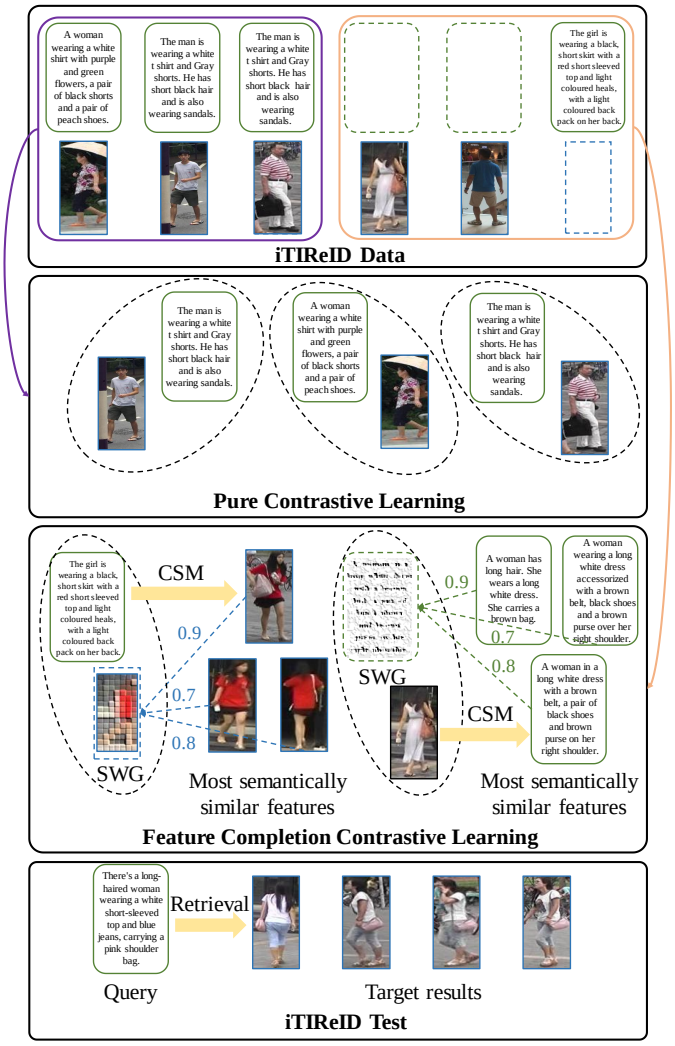

Fig. 1. Illustration of the proposed iTIReID task and two-stage contrastive completing Learning framework. iTIReID contains both complete pairwise data as well as incomplete data, and all data lacks identity labels. In pure contrastive learning stage, only complete paired data is used for training, which serves as a preliminary improvement of the model's capacity and prepares for the upcoming feature completion stage. In feature completion contrastive learning stage, the CSM leverages intra-modality similarity to measure cross-modal similarity and filter out the available features with the highest semantic similarity, thereby circumventing modality discrepancy. Subsequently, SWG generates approximate features based on the semantic similarity weight of the similar features. Contrastive CMPM loss, utilizing text-image pairs, performs contrastive learning in a weakly supervised manner.

and modality discrepancy cannot be effectively reduced. Existing methods can not perform well under these circumstances. We refer to this problem as the incomplete data problem. To ensure the robustness of the model in practical applications, it is important and necessary to address the issue of incomplete data. However, this problem has not been effectively solved in previous studies.

To enhance the practical scalability of TIReID in real-world scenarios, we must consider more practical settings. In this paper, we propose the task of incomplete text-image person re-identification (iTIReID) for the first time. Specifically, compared to the ideal conventional fully supervised TIReID, iTIReID has two practical and challenging settings. Firstly, to reduce the cost of data labeling, there are no sample identity labels in iTIReID, and only a few text-image pairs are available (see the iTIReID data in Fig. 1). In this setting, the common identity-based optimization strategy cannot be applied, and there is an urgent need to solve the problem of how to train and obtain a stable feature space that is reasonable. Secondly, most of the data in the iTIReID training set is incomplete. There are usually either images or text descriptions alone (see the iTIReID data in Fig. 1). This setting simulates two common real-world scenarios: 1) using a small amount of manually annotated text descriptions to reduce data acquisition cost and 2) scenarios where data is incomplete. In this setting, a significant amount of incomplete data cannot participate in cross-modality optimization. Addressing the challenge of how to achieve semantic alignment and reduce modality discrepancies using a limited number of complete pairwise and a large number of incomplete data is a very challenging problem.

We try to propose a solution approach for iTIReID. Intuitively, iTIReID can be divided into two sub-problems that are easily solvable: 1)completing incomplete modality features to ensure that features used in training are all text-image pairs, achieving modality balance; 2)using the corresponding relationships between text-image pairs as weak supervision signals for weakly supervised training.

Based on this motivation, we propose a Contrastive Completing Learning (CCL) framework to address iTIReID, which includes two learning stages (see Fig. 1). The first stage is the Pure Contrastive Learning (PCL) stage. Only complete pairwise data is used for training in this stage, which serves as a preliminary improvement of the model's capacity and prepares for the upcoming feature completion stage. To ensure fine-grained semantic alignment and reduce modality discrepancies, we propose a Prototype-based Cross-Modality Encoder (PCME) that maps image and text modalities to a common latent space, achieving implicit local semantic alignment. To fully utilize text-image pairs and facilitate weakly supervised training, we propose a contrastive CMPM (CCMPM) loss that maximizes the similarity of matching image-text pairs while minimizing the similarity of non-matching image-text pairs, thus obtaining a reasonable feature distribution where only pairwise samples have absolute high similarity in a contrastive learning manner.

The second stage is the Feature Completion Contrastive Learning (FCCL) stage. In this stage, available aligned features are proposed to complete missing modality features, enabling effective training with incomplete data. For example, for an incomplete image data, although there is no corresponding text description, there must be highly relevant descriptions that are semantically similar to it. These descriptions may come from images with the same identity as the incomplete image or from other images with extremely similar appearances. These descriptions can partially replace the missing text description to a certain extent. Therefore, we utilize the existing incomplete features to query the available complete features of the other modality, filter out the most semantically similar features, and complete the missing modality features based on semantic correlation. To achieve this, a Cross-modal Semantic Measure (CSM) and a Semantic-Weighted Generation (SWG) method

are proposed. The CSM leverages intra-modality similarity to measure cross-modal similarity and filter out features with the highest semantic similarity, thereby circumventing modality discrepancy. It can be understood that these features have a part of semantic content similar to the missing feature, with higher similarity indicating a greater amount of related semantic content. Subsequently, SWG generates approximate features based on the semantic similarity weight of the similar features. After feature completion, balance between different modality features is achieved, and we also adopt the contrastive CMPM loss to optimize the completed features. Experimental results show that the proposed framework effectively reduces modality discrepancy using incomplete data and achieves promising accuracy.

The main contributions of our work are as follows:

- We propose the incomplete text-image person re-identification (iTIReID) task for the first time, which considers more practical settings and aims to reduce the cost of data acquisition while improving the robustness of models. This has important implications for enhancing the scalability of TIReID technology in real-world scenarios.
- We propose a contrastive CMPM loss that maximizes the similarity of matching image-text pairs while minimizing the similarity of non-matching image-text pairs, thus fully utilizing text-image pairs to facilitate weakly supervised training.
- We introduce a cross-modal semantic measure and a semantic-weighted generation method to address incomplete data issues. In scenarios without identity labeling, CSM leverages intra-modality similarity to measure cross-modal similarity and filter out features with the highest semantic similarity, thereby circumventing modality discrepancy. SWG further generates approximate modality features based on semantic relevance, achieving balance between the features of two modalities.
- We propose a contrastive completing learning framework for iTIReID, which can be trained in a weakly supervised manner by only utilizing correspondence between samples. Additionally, CCL effectively addresses incomplete data issues, reducing the cost of data acquisition in practical applications while greatly enhancing model robustness.
- Our experimental results demonstrate that the proposed framework achieves promising accuracy in the iTIReID task. Compared with many fully supervised methods that use complete data, the performance of our proposed framework is highly competitive, demonstrating the superiority of our approach.

II. Related Work

A. Person Re-identification

Person re-identification aims to retrieve target individuals from a large number of pedestrian images or video clips collected in cross-camera scenarios. Compared to general retrieval tasks, person re-identification is challenging due to the large intra-class variability and small inter-class variability [1], [9]–[11]. The small inter-class variability arises from the fact that all the classes are pedestrians, while the large intra-class variability is caused by factors such as changes in camera angle, pedestrian posture, lighting conditions, and occlusion in shooting scenes. Research on person re-identification focuses primarily on two aspects: robust feature learning and effective metric learning.

Hermans et al. [12] proposed the use of triplet loss to optimize feature space by bringing positive samples closer together and push negative samples further apart to reduce intra-class variability and increase inter-class variability. Chen et al. [13] proposed a quadruplet deep network to enhance feature space optimization with a quadruplet loss. Sun et al. [14] proposed PCB, which adapted to mining fine-grained information by splitting pedestrian images. Wang et al. [15] proposed MGN for capturing pedestrian features at multiple scales to obtain more discriminative features. In recent years, significant progress has been made in person re-identification research, and many methods have achieved matching accuracy that is close to or even surpasses manual matching accuracy.

However, in practical application scenarios, pedestrian images or video clips may not always be available due to missing road monitoring or obstruction of the target individual, which limits the potential application of ReID technology. Nonetheless, target individuals can still be searched based on verbal descriptions provided by eyewitnesses, which is referred to as TIReID [3]. In practical application scenarios, language descriptions are often easier to obtain and more flexible to use than images, making TIReID research highly valuable. Currently, TIReID is receiving increasing attention from both academia and industry.

B. Text-image Person Re-identification

The aim of TIReID is to leverage textual descriptions for the retrieval of target pedestrians from a vast image gallery [3]. In light of the substantial modality discrepancy between the image and text modalities, earlier approaches have sought to project data from these divergent modalities into a shared feature space to facilitate comparison. Faghri et al. [16] introduced a ranking loss function to improve embedding optimization, which seeks to minimize the intra-class distance while maximizing the inter-class distance. Sarafianos et al. [17] utilized an adversarial learning approach to encourage the textual and visual features to be indistinguishable by a modality discriminator. Meanwhile, Zheng et al. [18] proposed an instance loss function that leverages classifiers shared across both modalities to achieve superior inter-modal alignment. Zhang et al. [19] proposed a cross-modal projection matching (CMPM) loss and a cross-modal projection classification (CMPC) loss for learning discriminative image-text embeddings. As the research progresses, studies on TIReID have increasingly focused on mining fine-grained features and semantic alignment. Chen et al. [20] proposed optimizing the affinity between images and text by leveraging the similarity between words and image patches. Niu et al. [21] introduced a cross-modal attention mechanism that aligns semantics at three different scales: global-to-global, global-to-local, and local-to-local. Wang et al. [5] utilized an auxiliary segmentation model

to extract key nouns and image patches, and aligned them accordingly. Shao et al. [6] proposed to learn granularity-unified representations for both text and image modalities. Jiang et al. [22] proposed SDM loss to minimizes the KL divergence between the normalized image-text similarity score distributions and the normalized ground truth label matching distributions.

However, these methods require a large number of images, detailed descriptions, and identity labels for training, incurring a high cost of data acquisition that is difficult to justify in practical applications. Additionally, existing methods assume that data from multiple modalities are complete and balanced, which is often difficult to satisfy in actual scenarios due to difficulties in data acquisition or improper storage that may lead to data loss. These two critical issues have hindered the further applications of TIReID in practical scenarios. Therefore, in this paper, we propose the iTIReID task to reduce data dependence and enhance robustness in actual application scenarios. Furthermore, we introduce a novel contrastive completing learning framework for iTIReID. CCL utilizes available features to complete missing features under weakly-supervised conditions and performs weakly-supervised contrastive learning to effectively utilize incomplete data and reduce modality differences.

C. Weakly Supervised Text-image Retrieval

To reduce TIReID's dependence on identity labels, some studies have explored effective training methods under weakly-supervised conditions. Patel et al. [23] leveraged the correlations between images and text found throughout a set of articles for training data, optimizing the common feature space at both global and local scales. Meanwhile, Gomez et al. [24] exploited freely available paired data to learn a multimodal image and text embedding. Zhao et al. [25] proposed CMMT to leverage pseudo labels for self-training in each modality and uses similarity soft-labels to facilitate cross-modal matching learning.

While these methods reduce TIReID's dependence on identity labels to a certain extent, they do not diminish the need for more expensive manually annotated text descriptions. Additionally, these methods are unable to effectively handle incomplete data in actual application scenarios. CCL utilizes available features to complete missing features under weakly-supervised conditions, while performing weakly-supervised contrastive learning training to effectively utilize incomplete data and reduce modality differences. Experimental results demonstrate the effectiveness of CCL in addressing iTIReID and achieving competitive accuracy in real-world application scenarios.

D. Modality-imbalance Research

In multimodal research, modality imbalance is a common problem that arises from missing features in certain modalities. To address this issue, most methods generate corresponding features to achieve modality balance and facilitate subsequent optimization. Wu et al. [26] utilized adversarial training to learn hash mappings between modalities and generate features. Guo et al. [27] proposed a collective affinity learning method and further introduced a probability model to reconstruct features. Some deep learning-based methods tend to use the centroid of related features as the semantic center or cluster center, which is simple and effective, and has strong stability in relation to changes in feature quantity. Zeng et al. [28] proposed using class centers as semantic centers to reconstruct features and predict invisible classes, while Liu et al. [29] used proxies of equivalent modalities to benefit cross-modality association.

However, in iTIReID, using the centroid as a substitute feature for missing features is not a reasonable approach. It is important to consider that approximate features have some semantic relevance to the missing ones, and the higher the similarity, the more relevant the semantics are. However, using centroids does not consider the semantic similarity weight. In this paper, we propose a cross-modal semantic measure and a semantic-weighted generation method. CSM utilizes intra-modal sample similarity to avoid modality differences and measure cross-modal similarity by selecting the most semantically similar features. SWG further generates approximate modality features based on semantic similarity weight. Experimental results confirm the effectiveness and superiority of the proposed methods.

III. Method

In this section, we will elaborate the proposed iTIReID task and the contrastive completing learning framework. First, we provide the problem definition of the iTIReID task in Section III-A. Then in Section III-B, we introduce the pure contrastive learning stage of CCL, which includes the proposed prototype-based cross-modality encoder and contrastive CMPM losses. In Section III-C, we discuss the feature completion contrastive learning stage of CCL, which includes the proposed cross-modal semantic measure and semantic-weighted generation. The overview of the proposed framework is illustrated in Fig. 2, and the overall training procedure is shown in Alg. 1.

A. Problem Definition

In fully supervised TIReID task, it is assumed that the data from each modality are balanced. For each image, there is a corresponding description in the text modality with the same identity label. However, in practical application scenarios, such ideal conditions are typically difficult to achieve due to the cost and difficulty of data acquisition. Therefore, we propose iTIReID, which aims to reduce the data cost required for model training and enhance the robustness of the model, strengthening the scalability of image-text person re-identification technology in practical scenarios.

For clarity, we use superscripts v and t to represent variables from the image and text modalities, respectively, and use superscript s to indicate variables associated with incomplete data. Without loss of generality, the training dataset in iTIReID comprises complete pairwise data denoted as $\{(x_i^v, x_i^t)\}_{i=1}^{N}$, where x_i^v represents the image and x_i^t represents the textual description, and incomplete data denoted as $\{x_i^{vs}\}_{i=N+1}^{N^v}$ and $\{x_i^{ts}\}_{i=N+1}^{N^t}$, representing instances where only one modality

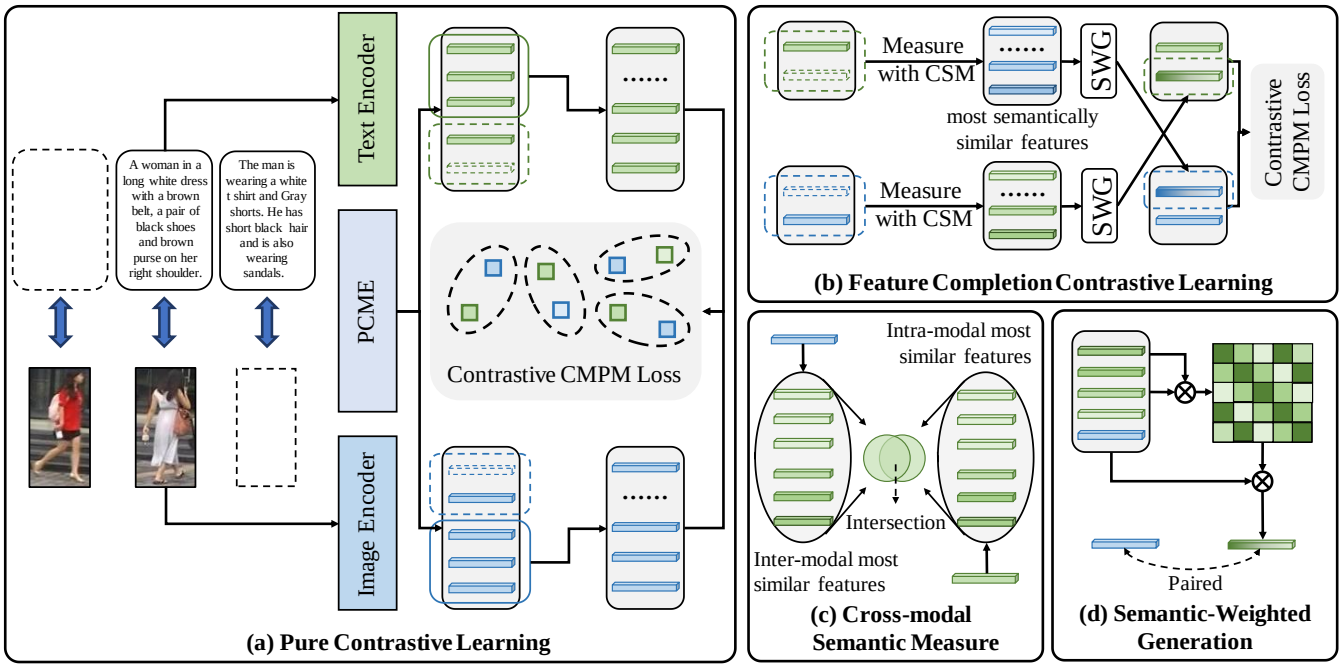

Fig. 2. The overview of the proposed contrastive completing learning framework (CCL), which includes a pure contrastive learning (PCL) stage and a feature completion contrastive learning (FCCL) stage. In the PCL stage, only complete paired data are adopted for training. The prototype-based cross-Modality encoder (PCME) performs implicit local semantic alignment and maps images and text modalities to a common latent space. In the FCCL stage, we propose utilizing available complete features to complete missing modality features, enabling effective training using incomplete data. The cross-modal semantic measure (CSM) leverages intra-modality sample similarity to measure cross-modal similarity and filter out features with the highest semantic similarity, thereby circumventing modality discrepancy. The semantic-weighted generation (SWG) generates approximate modality features based on the semantic similarity weight of the similar features. The contrastive CMPM (CCMPM) loss that fully utilizes text-image pairs to facilitate weakly supervised training.

is available and the corresponding data from the other modality is missing. N denotes the number of complete paired data, while N^v and N^t respectively denote the total number of images and textual descriptions. In the training dataset, identity labels are unavailable for all cases, and only text-image correspondence relationships can be utilized.

The main objective of iTIReID is to train a robust model that can retrieve target pedestrian images matching textual description from a large image database containing various pedestrian images.

B. Pure Contrastive Learning

In iTIReID, complete paired data with both images and accurately annotated text descriptions are a valuable training resource. During the pure contrastive learning stage, only complete paired data is utilized for training in order to reduce modal discrepancy and improve model performance.

Prototype-based Cross-Modality Encoder. In cross-modal tasks, we expect to map features from different modalities to a joint embedding space with sufficient modality interaction and semantic alignment at the feature-level. Conventional TIReID approaches typically rely on explicitly establishing correspondences between local regions for local matching. However, these methods are no longer suitable in iTIReID due to the lack of identity labels and missing data. Therefore, we propose a prototype-based cross-modality encoder for implicit local semantic alignment, reducing modality differences and facilitating distance measurement between features, as shown in Fig. 3.

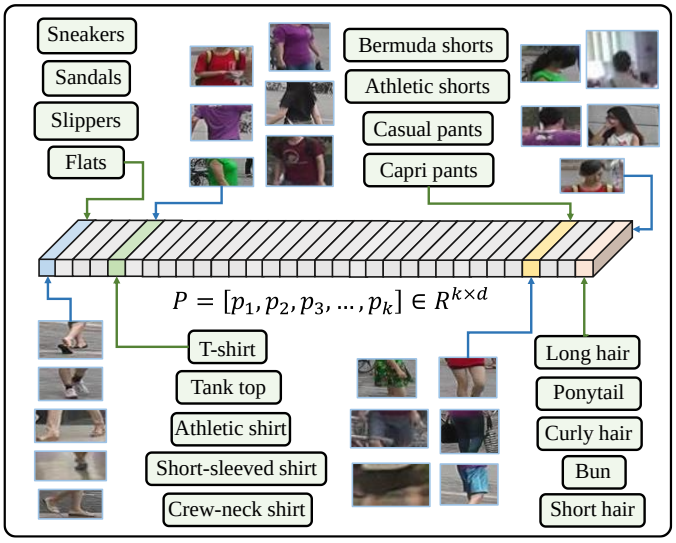

Fig. 3. Illustration of proposed PCME. PCME serves for implicit local semantic alignment, primarily aligning at the level of image patches and either words or phrases.

Specifically, we design a set of prototypes $P = [p_1, p_2, \ldots, p_k] \in \mathbb{R}^{k \times d}$ that are shared across modalities and randomly initialized. k denotes the number of prototypes, and d represents the dimension of features. We expect to capture features from the textual and image modalities simultaneously with the prototypes, achieving adaptive cross-modal semantic alignment. We utilize cross-attention operations [30]

Algorithm 1 Training Procedure of CCL.

Input: $\{(x_i^v, x_i^t)\}_{i=1}^{N}$: complete text-image pair; $\{x_i^{vs}\}_{i=N+1}^{N^v}$: incomplete images; $\{x_i^{ts}\}_{i=N+1}^{N^t}$: incomplete texts; $MaxEpochs\_PCL$: total training epochs for PCL; $MaxEpochs\_FCCL$: total training epochs for FCCL; lr: learning rate; Φ: learnable parameters of CCL.

Output: CCL_Φ with strong retrieval ability.

Pure Contrastive Learning:

for $epoch = 1$ **to** $MaxEpochs\_PCL$ **do**
  Obtain $\{(z_i^v, z_i^t)\}_{i=1}^{N}$ extracted from $\{(x_i^v, x_i^t)\}_{i=1}^{N}$ with the backbone network;
  Obtain $\{(f_i^v, f_i^t)\}_{i=1}^{N}$ from $\{(z_i^v, z_i^t)\}_{i=1}^{N}$ with PCME;
  Compute loss $\mathcal{L}_{ccmpm}$ on $\{(f_i^v, f_i^t)\}_{i=1}^{N}$;
  Update $\Phi' \leftarrow Adam(\bigtriangledown_\Phi \mathcal{L}_{ccmpm}, \Phi, lr)$;
end for

Feature Completion Contrastive Learning:

for $epoch = 1$ **to** $MaxEpochs\_FCCL$ **do**
  Obtain $\{(z_i^v, z_i^t)\}_{i=1}^{N}$ extracted from $\{(x_i^v, x_i^t)\}_{i=1}^{N}$ with the backbone network;
  Obtain $\{(f_i^v, f_i^t)\}_{i=1}^{N}$ from $\{(z_i^v, z_i^t)\}_{i=1}^{N}$ with PCME;
  for m **in** (v, t) **do**
    Obtain $\{z_i^{ms}\}_{i=N+1}^{N^m}$ extracted from $\{x_i^{ms}\}_{i=N+1}^{N^m}$ with the backbone network;
    Obtain $\{f_i^{ms}\}_{i=N+1}^{N^m}$ from $\{z_i^{ms}\}_{i=N+1}^{N^m}$ with PCME;
    Obtain $\{\mathcal{N}(f_i^{ms}, k_{vs}, T_{inter})\}_{i=N+1}^{N^m}$ with CSM;
  end for
  Obtain $\{f_i^{tg}\}_{i=N+1}^{N^v}$ with $\{\mathcal{N}(f_i^{vs}, k_{vs}, T_{inter})\}_{i=N+1}^{N^v}$ via SWG;
  Obtain $\{f_i^{vg}\}_{i=N+1}^{N^t}$ with $\{\mathcal{N}(f_i^{ts}, k_{vs}, T_{inter})\}_{i=N+1}^{N^t}$ via SWG;
  Compute loss $\mathcal{L}_{ccmpm}^{s}$ on $\{(f_i^{vs}, f_i^{tg})\}_{i=N+1}^{N^v}$ and $\{(f_i^{vg}, f_i^{ts})\}_{i=N+1}^{N^t}$;
  Update $\Phi' \leftarrow Adam(\bigtriangledown_\Phi \mathcal{L}_{ccmpm}^{s}, \Phi, lr)$;
end for

from transformer to achieve this procedure. Let $\{z_i^v\}_{i=1}^{N}$ and $\{z_i^t\}_{i=1}^{N}$ denote the features of complete paired data output by the backbone network. The output features serve as key and value while P serves as query. This process can be formulated as follows:

$$f_i^v = MHA(MCA(P, z_i^v, z_i^v)), \tag{1}$$

$$f_i^t = MHA(MCA(P, z_i^t, z_i^t)), \tag{2}$$

where MCA denotes the multi-head cross attention and MHA denotes a transformer block, which includes a multi-head attention and a feed-forward network. f_i^v and f_i^t are the feature embeddings output by PCME. After the cross-attention process with modality-shared prototype P, features from different modalities have been mapped to the aligned joint embedding space. All complete paired data features output from PCME can be denoted as $\{(f_i^v, f_i^t)\}_{i=1}^{N}$. Similarly, we can obtain the embedded incomplete data features as $\{f_i^{vs}\}_{i=N+1}^{N^v}$ and $\{f_i^{ts}\}_{i=N+1}^{N^t}$. In the Pure Contrastive Learning stage, only the complete data features are involved in training.

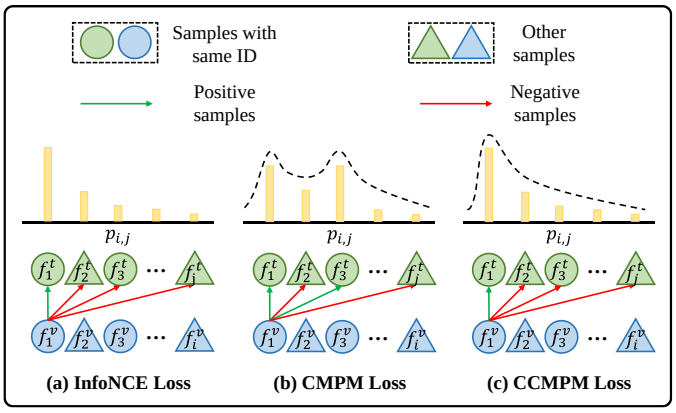

Fig. 4. Illustration of infoNCE loss, CMPM loss and CCMPM loss. The blue shapes represent the image modality, while the green shapes represent the text modality. The CMPM loss utilizes identity labels, treating all samples with the same identity as positive samples, aiming to obtain a feature distribution that matches the label distribution. The infoNCE loss only considers corresponding samples as positive samples, aiming to maximize the similarity between positive samples and minimize the similarity between negative samples. The CCMPM loss does not use identity labels and treats pairwise samples as positive samples. It combines the advantages of the previous two losses, aiming to obtain a feature distribution where only positive samples have absolute high similarity.

Contrastive CMPM Loss. In fully supervised TIReID, identity optimization is highly effective and stable, achieving satisfactory accuracy. However, in iTIReID, identity optimization is not applicable due to the lack of identity labels. Label-free ReID methods tend to utilize clustering to obtain pseudo-labels, enabling identity optimization in the feature space. In contrast to these methods, iTIReID involves valuable correspondences between image and text descriptions that can be utilized for optimization. CMPM loss [19] performs well on guiding image-text matching and shows stability with different batch sizes, but it requires identity labels. To explore more effective label-free cross-modal matching objectives, we propose a contrastive CMPM loss inspired by the training methods and InfoNCE loss [31] in visual-language pre-training (VLP) tasks [32], [33]. The comparison of three kinds of losses is illustrated in Fig. 4. CCMPM loss maximizes the similarity of matching image-text pairs while minimizing the similarity of non-matching image-text pairs, thus obtaining a reasonable feature distribution where only pairwise samples have absolute high similarity.

In a setup similar to contrastive learning, we considers paired images and text descriptions as positive samples, while unpaired images and text descriptions are considered negative samples. Specifically, for a pair of image and text description (f_i^v, f_j^t), if $i = j$, they are considered positive samples; if $i \neq j$, they are considered negative samples. The probability of image-text matching can be formulated as follows:

$$p_{i,j} = \frac{\exp(\overline{f_i^v}\ \overline{f_j^t}/\tau)}{\sum_{k=1}^{M}\exp(\overline{f_i^v}\ \overline{f_k^t}/\tau)} \quad s.t. \quad \overline{f} = \frac{f}{\|f\|}, \tag{3}$$

where τ is a temperature coefficient and M is the size of a mini-batch, and $M < N$. $p_{i,j}$ can be understood as the

proportion of similarity between f_i^v and f_j^t that is accounted for in the similarity between f_i^v and the current batch $\{f_k^t\}_{k=1}^M$. Following the approach of contrastive learning, we aim to ensure that when f_i^v and f_j^t are positive samples, their similarity is sufficiently high, and when they are negative samples, their similarity is sufficiently low. The contrastive loss from image to text can then be formulated as:

$$\mathcal{L}_i = \sum_{j=1}^{M} p_{i,j} \log \frac{p_{i,j}}{\mu + \epsilon} \quad s.t. \quad \mu = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases}, \tag{4}$$

$$\mathcal{L}_{v2t} = \frac{1}{M} \sum_{i=1}^{M} \mathcal{L}_i, \tag{5}$$

where ϵ is a small number to avoid computational issues. Similarly, the contrastive loss from text to image can then be formulated by exchanging f^v and f^t in Eq. (3) (4) (5), and the bi-directional contrastive CMPM loss is calculated by:

$$\mathcal{L}_{ccmpm} = \mathcal{L}_{v2t} + \mathcal{L}_{t2v} \tag{6}$$

C. Feature Completion Contrastive Learning

In iTIReID, the majority of the data consists of incomplete entries that typically only contain either images or text descriptions, with only a small amount of complete paired data available. During the feature completion stage of contrastive learning, we propose the use of available features to fill in missing modality features, achieving modality balance and effective utilization of incomplete data for training. For ease of description, in the following discussion we use the process of completing incomplete text features as an example, and the same process can be applied to incomplete image features.

Cross-modal Semantic Measure. Reducing modality discrepancy is a critical target of multi-modal tasks. Many studies have focused on achieving modality balance through feature generation, which in turn helps reduce modality discrepancy [34]–[36]. This inspired us to address data incompleteness and modality imbalance in iTIReID. For an incomplete text feature $f_i^{ts} \in \{f_i^{ts}\}_{i=N+1}^{N^t}$, although it does not have a corresponding image, there must be images that are highly semantically similar to it. These images may come from images belonging to the same identity as f_i^{ts}, or from images with extremely similar appearances. Therefore, these images can substitute for the missing image to some extent. We propose the use of similar features to complete the originally incomplete data and achieve training in a modality-balanced state.

Specifically, for an text feature f_i^{ts}, we need to measure the similarity between f_i^{ts} and all complete data images $\{f_j^v\}_{j=1}^N$, with higher similarities indicating greater semantic relevance. Intuitively, we might think of naive cosine similarity or Euclidean distance to measure this. However, searching for similar features with f_i^{ts} is essentially a cross-modal matching process, and we cannot guarantee accuracy because of modality discrepancy. As such, we propose cross-modal semantic measure, which leverages intra-modality sample similarity to circumvent modality discrepancy when measuring cross-modal similarity, in order to select the most semantically similar features. The process of CSM is illustrated in Fig. 5.

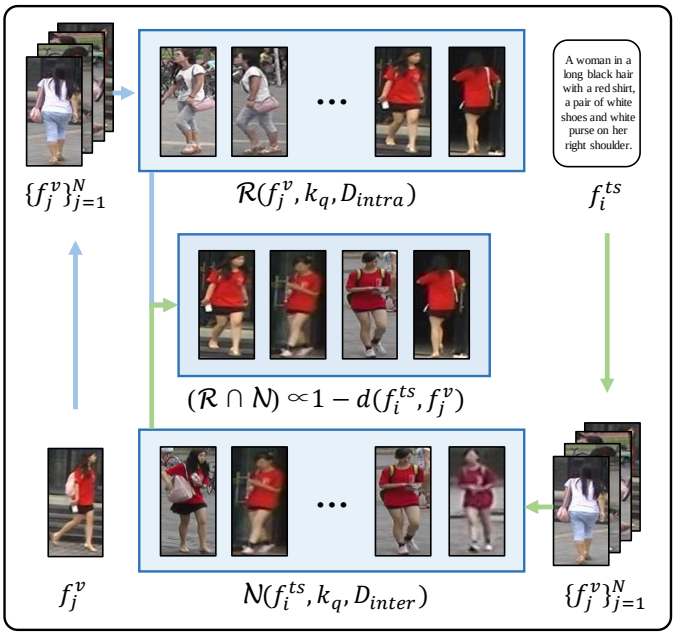

Fig. 5. Illustration of proposed CSM. The acquisition process of $\mathcal{R}(f_j^v, k_q, D_{intra})$ has been simplified. f_i^{ts} and f_j^v query $\left\{f_j^v\right\}_{j=1}^N$, to obtain semantically most similar neighbors, $\mathcal{N}(f_i^{ts}, k_q, D_{inter})$ and $\mathcal{R}(f_j^v, k_q, D_{intra})$, respectively. According to Eq. (9), $(\mathcal{R} \cap \mathcal{N})$ and $1 - d(f_i^{ts}, f_j^v)$ have a direct proportionate relationship.

For an $f_i^{ts} \in \{f_i^{ts}\}_{i=N+1}^{N^t}$ and an $f_j^v \in \left\{f_j^v\right\}_{j=1}^N$, if they both have the same mutual k semantically most similar neighbors in $\left\{f_j^v\right\}_{j=1}^N$, it indicates that they are more semantically related. Similarly, this measure incorporating neighbor information has been explored in re-ranking methods [37]. Re-ranking is typically used in the post-processing of single-modal ReID retrieval to improve final result accuracy, but it cannot handle modality discrepancy well. Our CSM leverages intra-modality similarity to circumvent modality discrepancy and measures cross-modal similarity to select the most similar cross-modal samples to assist training. Given an f_i^{ts}, we rank its correlation with $\left\{f_j^v\right\}_{j=1}^N$ using D_{inter} and obtain its k_q semantically most similar cross-modal neighbors:

$$\mathcal{N}(f_i^{ts}, k_q, D_{inter}) = \left\{f_{(1)}^v, f_{(2)}^v, \ldots, f_{(k_q)}^v\right\}, \tag{7}$$

where D_{inter} is the cosine similarity matrix between f_i^{ts} and $\left\{f_j^v\right\}_{j=1}^N$. Given an f_j^v, we rank its correlation with $\{f_k^v\}_{k=1}^N$ using D_{intra} and obtain its k_q semantically most similar reciprocal neighbors within the modality as:

$$\mathcal{R}(f_j^v, k_q, D_{intra}) = \mathcal{N}(f_j^v, k_q, D_{intra}) \cap \mathcal{N}(f_k^v, k_q, D_{intra}), \tag{8}$$

where D_{intra} is the cosine similarity matrix between f_j^v and $\{f_k^v\}_{k=1}^N$. The mutual k semantically most similar neighbors of f_i^{ts} and f_j^v is the intersection of $\mathcal{N}(f_i^{ts}, k_q, D_{inter})$ and $\mathcal{R}(f_j^v, k_q, D_{intra})$. Therefore, semantic distance between f_i^{ts}

and f_j^v can be defined as:

$$d(f_i^{ts}, f_j^v) = 1 - \frac{\left|\mathcal{N}(f_i^{ts}, k_q, D_{inter}) \cap \mathcal{R}(f_j^v, k_q, D_{intra})\right|}{\left|\mathcal{N}(f_i^{ts}, k_q, D_{inter}) \cup \mathcal{R}(f_j^v, k_q, D_{intra})\right|}, \tag{9}$$

where $|\cdot|$ denotes the number of elements in the set. Obviously, the more mutual k semantically most similar neighbors $\mathcal{N}(f_i^{ts}, k_q, D_{inter})$ and $\mathcal{R}(f_j^v, k_q, D_{intra})$ have, the higher the semantic relevance between f_i^{ts} and f_j^v, and the smaller distance. We use CSM to measure and obtain the k_{vs} most semantically relevant features for f_i^{ts} in $\left\{f_j^v\right\}_{j=1}^{N}$ as:

$$\mathcal{N}(f_i^{ts}, k_{vs}, T_{inter}) = \left\{f_{(1)}^v, f_{(2)}^v, \ldots, f_{(k_{vs})}^v\right\}, \tag{10}$$

where T_{inter} is the semantic similarity matrix between f_i^{ts} and $\left\{f_j^v\right\}_{j=1}^{N}$, measured with CSM.

Semantic-Weighted Generation. The k_{vs} most semantically relevant features are obtained and used to generate missing image modality features. A naive approach would be to directly use the most similar feature as a replacement for the missing feature. However, even the most semantically similar features generally only contain a part of the missing feature's semantic content, making this approach overly simplistic. In past cross-modal retrieval research, the centroid of relevant features has been used as the semantic or cluster center [28], [29]. This method is simple and effective, exhibiting strong stability across different feature numbers. However, in iTIReID, using the centroid as a replacement for the missing feature is not a reasonable approach. These features may have some semantic relevance to the missing feature, with higher similarity indicating a greater amount of related semantic content. Using the centroid without considering semantic similarity weight would be inaccurate. To address this issue, we propose a semantic-weighted generation method, as illustrated in Fig. 6, which generates approximate features based on semantic similarity weight. This process can be formulated as follows:

$$\begin{aligned} E &= [e_1, e_2, \ldots, e_{k_{vs}+1}] \\ &= \left[W(f_i^{ts}), W(f_{(1)}^v), W(f_{(2)}^v), \ldots, W(f_{(k_{vs})}^v)\right], \end{aligned} \tag{11}$$

$$f_i^{vg} = A_{rlt} \cdot E, \tag{12}$$

where f_i^{vg} is the generated image feature, and the superscript g indicates a variable related to the generated feature. W is a non-linear transformation with learnable parameters. A_{rlt} represents the affinity matrix, which represents the semantic correlation between f_i^{ts} and other relevant features. A_{rlt} is computed from the similarity matrix, which can be formulated as follows:

$$S_{i,j} = \exp(\overline{e_i}\ \overline{e_j}) \quad s.t. \quad \overline{e} = \frac{e}{\|e\|}, \tag{13}$$

$$A_{rlt} = D^{-1} \cdot S, \tag{14}$$

where D^{-1} is the Laplacian matrix of S, which is used to normalize S. Thus, we have completed the feature completion process. By completing $\{f_i^{ts}\}_{i=N+1}^{N^t}$, we obtain $\left\{f_j^{vg}\right\}_{j=N+1}^{N^t}$. Similarly, by completing $\{f_i^{vs}\}_{i=N+1}^{N^v}$, we can obtain $\left\{f_j^{tg}\right\}_{j=N+1}^{N^v}$. Like complete paired features, for a pair of features $\left(f_i^{vs}, f_j^{tg}\right)$ or $\left(f_i^{ts}, f_j^{vg}\right)$, if $i = j$, they are considered positive samples; if $i \neq j$, they are considered negative samples. We also adopt contrastive CMPM loss for optimization:

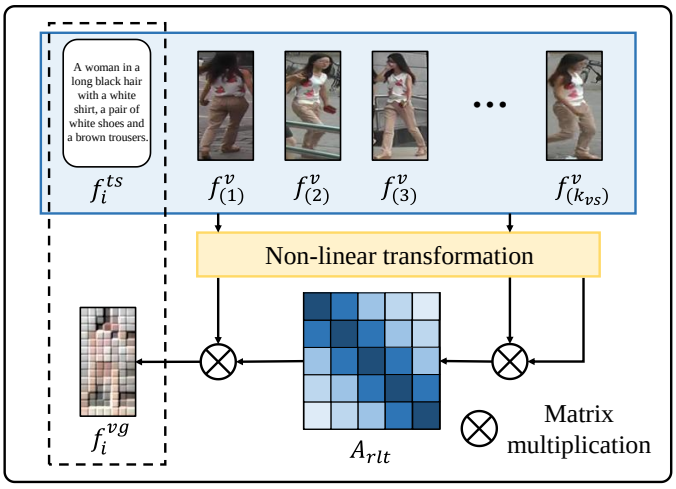

Fig. 6. Illustration of proposed SWG. The dashed box indicates that f_i^{ts} and f_i^{vg} positive samples.

$$\mathcal{L}_{ccmpm}^s = \mathcal{L}_{v2t}^s + \mathcal{L}_{t2v}^s, \tag{15}$$

IV. Experiments

A. Experimental Setup

CUHK-PEDES [3] is the first large-scale publicly available dataset for the TIReID task. It consists of 13003 pedestrians with 40206 images and corresponding 80412 manually annotated descriptions, where each image has two hand-crafted descriptions and the average length of each description is no less than 23 words. Specifically, 11003 pedestrians with 34054 images and 68108 descriptions are used as training set, 1000 pedestrians with 3078 images and 6156 descriptions are used as validation set, and another 1000 pedestrians with 3074 images and 6148 descriptions are used as testing set.

ICFG-PEDES [7] contains 54522 images and corresponding 54522 descriptions of 4102 pedestrians, which are selected from the MSMT17 dataset. The statistics show that the average length of descriptions in ICFG-PEDES is 37 words and there are 5554 unique vocabulary words in all the descriptions. Compared with CUHK-PEDES, the samples in ICFG-PEDES are more fine-grained and identity-centric. Specifically, 34674 sample pairs of 3102 pedestrians and 19848 sample pairs of the remaining 1000 pedestrians are respectively used as training and testing sets.

RSTPReid [38] comprises 20,505 images and corresponding 41,010 descriptions of 4,101 pedestrians. Each identity has five images, each with two corresponding text descriptions, with each sentence containing no fewer than 23 words. All images are captured by 15 different cameras. For data partitioning, 3,701, 200, and 200 identities are used for training, validation, and testing, respectively.

TABLE I
PERFORMANCE COMPARISONS ON CUHK-PEDES UNDER FULL DATA MODE.

Methods	ID	Feat	Ref	Rank-1	Rank-5	Rank-10	mAP
Dual Path [18]	✓	G	TOMM20	44.40	66.26	75.07	-
CMPM/C [19]	✓	G	ECCV18	49.37	-	79.27	-
MIA [21]	✓	L	TIP20	53.10	75.00	82.90	-
ViTAA [5]	✓	L	ECCV20	55.97	75.84	83.52	51.60
IMG-Net [39]	✓	L	JEI20	56.48	76.89	85.01	-
SUM [40]	✓	L	KBS22	59.22	80.35	87.60	-
DSSL [38]	✓	G	MM21	59.98	80.41	87.56	-
SSAN [7]	✓	L	arXiv21	61.37	80.15	86.73	-
ISANet [41]	✓	L	arXiv22	63.92	82.15	87.69	-
IVT [42]	✓	G	ECCVW22	64.00	82.72	88.95	-
LBUL [43]	✓	L	MM22	64.04	82.66	87.22	-
TextReID [44]	✓	G	BMVC21	64.08	81.73	88.19	60.08
LGUR [6]	✓	L	MM22	64.21	81.94	87.93	-
TIPCB [45]	✓	L	Neuro22	64.26	83.19	89.10	-
CAIBC [46]	✓	L	MM22	64.43	82.87	88.37	-
AXM-Net [47]	✓	L	AAAI22	64.44	80.52	86.77	58.73
MM-TIM [24]	×	G	MSU2019	45.35	63.78	70.63	-
CMMT [25]	×	G	ICCV21	57.10	78.14	85.23	-
CCL(ours)	×	G	-	**67.25**	**86.10**	**91.45**	**60.83**

TABLE II
PERFORMANCE COMPARISONS ON ICFG-PEDES UNDER FULL DATA MODE.

Methods	ID	Feat	Ref	Rank-1	Rank-5	Rank-10	mAP
Dual Path [18]	✓	G	TOMM20	38.99	59.44	68.41	-
CMPM/C [19]	✓	G	ECCV18	43.51	65.44	74.26	-
MIA [21]	✓	L	TIP20	46.49	67.14	75.18	-
SCAN [8]	✓	L	ECCV18	50.05	69.65	77.21	-
ViTAA [5]	✓	L	ECCV20	50.98	68.79	75.78	-
SSAN [7]	✓	L	arXiv21	54.23	72.63	79.53	-
TIPCB [45]	✓	L	Neuro22	54.96	74.72	81.89	-
IVT [42]	✓	G	ECCVW22	56.04	73.60	80.22	-
ISANet [41]	✓	L	arXiv22	57.73	75.42	81.72	-
CCL(ours)	×	G	-	**58.33**	**76.75**	**83.38**	**32.66**

Training mode. In the iTIReID task, the training phase does not involve identity labels and some data may be incomplete, while the testing phase is the same as the TIReID task. To fully demonstrate the performance of the proposed framework for iTIReID, we designed multiple training modes, including full data mode, incomplete data mode, and incomplete text mode. The full data mode adopts all data for training, demonstrating the upper limit of the model's performance. The incomplete data mode simulates the most realistic application scenario, where the incomplete data used for training includes both images and texts. The incomplete text mode is designed to show the model's performance when trained with only a small amount of text descriptions. To achieve training with incomplete data, we partition the existing datasets by maintaining the testing set and partitioning only the training set. Specifically, we provide three settings of difficulty. For the incomplete data mode, under the easy setting, 50% of the training data are complete data, 25% of the training data are only image data, and 25% of the training data are only text data. The proportion of incomplete data increases under the medium and hard setting, with corresponding partition ratios of (30%, 35%, 35%) and (10%, 45%, 45%). For the incomplete text mode, the missing data are all text descriptions, so under the easy setting, 50% of the training data are complete data and 50% of the training data are only image data. Under the medium and hard setting, there is more missing texts, with corresponding partition ratios of (30%, 70%) and (10%, 90%).

TABLE III
PERFORMANCE COMPARISONS ON RSTPREID UNDER FULL DATA MODE.

Methods	ID	Feat	Ref	Rank-1	Rank-5	Rank-10	mAP
IMG-Net [39]	✓	L	JEI20	37.60	61.15	73.55	-
AMEN [48]	✓	G	PRCV21	38.45	62.40	73.80	-
DSSL [38]	✓	G	MM21	39.05	62.60	73.95	-
SUM [40]	✓	L	KBS22	41.38	67.48	76.48	-
SSAN [7]	✓	L	arXiv21	43.50	67.80	77.15	-
LBUL [43]	✓	L	MM22	45.55	68.20	77.85	-
IVT [42]	✓	G	ECCVW22	46.70	70.00	78.80	-
CCL(ours)	×	G	-	**51.30**	**75.25**	**84.60**	**41.10**

Evaluation Metrics. We employ recall at Rank K (Rank-K, higher is better) as primary retrieval metric to evaluate the retrieval performance. Given a textual description as query, if at least one matching image in the top-k candidate list, we call this an efficacious search. Moreover, we also employ mean Average Precision (mAP, higher is better) as a supplementary evaluation metric for more comprehensive evaluation. For all experiments, Rank-1, Rank-5, Rank-10 and mAP are reported.

Implementation details. In our experiments, we adopt the image encoder and text encoder of the Clip [32] model as the image and text feature extractors for our backbone network. During training, we use random horizontal flipping, random cropping, and random erasing as data augmentation methods for the image modality, with an input image size of 384*128. For the text modality, we set the maximum token sequence length to 80. We use the Adam optimizer [49] for optimization. For pure contrastive learning, we adopt an initial learning rate of 1e-5 and train for 60 epochs, with the learning rate reduced by a factor of 0.1 at the 20th and 50th epoch. Next, we proceed to feature completion contrastive learning, where we train for another 60 epochs with the same learning rate setting as PCL. Our batch size is set to 64. k is set as 500 for PCME, and τ is set as 0.02 for the CCMPM loss. k_q is set as 7 for CSM and k_{vs} is set as 5 for SWG.

B. Results and Comparisons

Full data mode. To validate the effectiveness of the proposed framework on TIReID task and demonstrate its general performance, we first compared CCL with existing methods on the CUHK-PEDES, ICFG-PEDES and RSTPReid datasets under the full data mode, as shown in Table I, Table II and Table III, respectively. Most of the methods participating in the comparison are fully supervised methods, including TextReID [45], which uses transfer learning to effectively transfer the Clip model's image-text matching ability to fine-grained TIReID to cope with the lack of large-scale datasets. LGUR [6] proposes to align images and texts at a granular level for effective semantic alignment, while AXM-Net [47] proposes a unified multi-layer network that can dynamically mine and align fine-grained information. A few methods are weakly supervised, including CMMT [25], which uses relationship between image-text pairs to reduce the noise in the generated

TABLE IV
PERFORMANCE COMPARISONS ON CUHK-PEDES UNDER INCOMPLETE DATA MODE. RESULTS OF EASY, MEDIUM AND HARD SETTINGS ARE REPORTED.

ID	Methods	Easy Setting(50%, 25%, 25%)				Medium Setting(30%, 35%, 35%)				Hard Setting(10%, 45%, 45%)			
		Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10	mAP
✓	CMPM/C [19]	40.65	62.96	74.63	35.15	40.09	62.45	74.93	35.56	27.04	52.62	64.03	25.12
✓	MIA [21]	46.17	68.49	77.73	39.29	43.58	66.41	73.96	37.00	30.14	55.12	65.69	26.83
✓	SCAN [8]	50.06	72.06	79.46	42.80	47.16	69.21	77.57	40.62	31.94	53.79	65.10	27.74
✓	ViTAA [5]	49.31	70.77	79.43	42.16	47.21	70.01	78.76	40.17	31.31	55.03	64.89	27.34
✓	TIPCB [45]	58.79	80.09	87.15	50.69	52.03	74.73	82.88	44.75	21.76	43.98	54.83	19.14
✓	SSAN [7]	53.32	74.43	82.22	45.31	48.96	71.83	79.62	38.81	34.11	57.92	68.21	29.71
✓	LGUR [6]	57.67	78.06	85.41	49.12	52.62	73.64	81.15	45.75	31.28	53.38	64.58	28.11
×	CCL (w/o FCCL)	60.05	81.22	88.23	54.73	50.55	75.04	82.74	44.23	29.02	54.41	65.35	28.20
×	CCL	**64.76**	**84.82**	**89.09**	**58.81**	**62.12**	**82.47**	**89.33**	**56.61**	**54.36**	**76.52**	**86.01**	**49.79**

TABLE V
PERFORMANCE COMPARISONS ON ICFG-PEDES UNDER INCOMPLETE DATA MODE. RESULTS OF EASY, MEDIUM AND HARD SETTINGS ARE REPORTED.

ID	Methods	Easy Setting(50%, 25%, 25%)				Medium Setting(30%, 35%, 35%)				Hard Setting(10%, 45%, 45%)			
		Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10	mAP
✓	CMPM/C [19]	34.68	55.88	65.53	17.41	29.64	51.15	60.83	15.15	15.50	30.13	41.46	9.06
✓	MIA [21]	38.93	60.24	69.17	20.09	35.81	58.38	67.71	18.60	19.66	35.40	45.96	10.50
✓	SCAN [8]	42.43	65.38	72.08	22.12	41.17	63.86	71.70	20.98	22.21	40.28	51.56	11.07
✓	ViTAA [5]	43.89	65.67	73.33	23.40	42.58	64.40	72.34	22.34	22.45	40.58	51.44	11.54
✓	TIPCB [45]	50.28	70.61	77.85	26.56	48.05	66.86	76.31	25.35	28.30	49.51	59.18	14.31
✓	SSAN [7]	46.36	66.87	75.40	24.11	44.85	66.70	74.64	23.91	24.92	42.46	53.01	13.01
✓	LGUR [6]	52.60	70.11	78.44	28.28	48.27	68.61	76.96	25.89	29.85	51.10	59.97	15.09
×	CCL (w/o FCCL)	50.08	70.70	78.25	26.97	43.98	64.97	73.47	23.03	14.13	30.11	39.35	8.28
×	CCL	**53.64**	**72.91**	**80.01**	**29.70**	**51.78**	**71.47**	**79.14**	**28.25**	**42.19**	**63.51**	**71.82**	**21.95**

pseudo labels of clustering. These methods had achieved state-of-the-art performance at the time.

CCL starts from more practical application conditions, and trains in a weakly supervised manner without using identity labels. As can be seen from Table I, our CCL achieves highly competitive results in CUHK-PEDES. Compared with fully supervised methods, CCL achieves the best results, achieving 67.25% on Rank-1, 86.10% on Rank-5, 91.45% on Rank-10 and 60.83% mAP respectively, surpassing AXM-Net Rank-1 by 2.81% , Rank-5 by 5.58%, Rank-10 by 4.68% and mAP by 2.1%. Compared with weakly supervised methods, CCL significantly outperforms CMMT, surpassing Rank-1 by 10.15%, Rank-5 by 7.96%, and Rank-10 by 6.22%, respectively. In ICFG-PEDES and RSTPReid, CCL achieves best results consistently. As shown in Table II, in ICFG-PEDES, CCL obtains 58.33% on Rank-1, 76.75% on Rank-5, 83.38% on Rank-10 and 32.66% on mAP. In RSTPReid, as shown in Table III, CCL obtains 51.30% on Rank-1, 75.25% on Rank-5, 84.60% on Rank-10 and 41.10% on mAP. This demonstrates that CCL has superior performance on the TIReID task and can learn a discriminative shared image-text feature space without using identity labels.

Incomplete data mode. To demonstrate the effectiveness of the proposed framework under more practical conditions, we test CCL on the CUHK-PEDES, ICFG-PEDES and RSTPReid under the incomplete data mode. To more intuitively demonstrate the impact of incomplete data, we select CMPM/C [19], MIA [21], SCAN [8], ViTAA [5], TIPCB [45], SSAN [7] and LGUR [6], which have been open-sourced, and compare them with the proposed framework in the incomplete data mode. Since previous methods did not focus on incomplete data, they are trained only using complete data. The results of these methods under incomplete data mode are obtained by running them ourselves. The results are shown in Table IV, Table V and Table VI. From the comparison of the results in the table, it can be observed that the performances of other methods have decreased significantly when the training data is incomplete, because these methods cannot effectively utilize these incomplete data for training. 'CCL (w/o FCCL)' refers to CCL with only PCL, meaning that CCL training uses only complete paired data and does not use incomplete data. In this case, the performance of CCL is lower than some of the comparative methods. CCL completes the incomplete data features by using semantically related features, thus can effectively utilize incomplete data to reduce modality discrepancy. Comparing 'CCL (w/o FCCL)' with 'CCL' in the Table IV, Table V and Table VI, after the feature completion contrastive learning stage, the performance of CCL has been greatly improved, demonstrating the rationality of our adopted strategy and the effectiveness of the completion method, and verifying the robustness of CCL to cope with iTIReID task. Under the most difficult setting of CUHK-PEDES, CCL achieves 54.36% on Rank-1, 76.52% on Rank-5, 86.01% on Rank-10 and 49.79% on mAP. Similarly, under the most difficult setting of ICFG-PEDES and RSTPReid, CCL achieves 42.19% and 42.15% on Rank-1, 63.51% and 68.10% on Rank-5, 71.82% and 78.15% on Rank-10, as well as 21.95% and 34.47% on mAP, significantly surpassing previous approaches. All of these results strongly demonstrate that CCL can effectively cope with practical application environments.

Incomplete text mode. To demonstrate the effectiveness of the proposed framework in reducing the amount of manually annotated text descriptions, we test CCL on the CUHK-PEDES, ICFG-PEDES and RSTPReid under the incomplete

TABLE VI
PERFORMANCE COMPARISONS ON RSTPREID UNDER INCOMPLETE DATA MODE. RESULTS OF EASY, MEDIUM AND HARD SETTINGS ARE REPORTED.

ID	Methods	Easy Setting(50%, 25%, 25%)				Medium Setting(30%, 35%, 35%)				Hard Setting(10%, 45%, 45%)			
		Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10	mAP
✓	CMPM/C [19]	32.82	56.87	67.50	27.52	29.51	54.13	65.14	25.92	17.07	38.29	48.70	17.22
✓	MIA [21]	37.62	60.50	71.50	31.35	35.11	58.22	70.90	29.57	21.43	43.08	55.19	20.61
✓	SCAN [8]	40.30	63.30	74.65	32.89	37.94	61.45	73.81	31.44	25.61	46.96	58.07	22.91
✓	ViTAA [5]	40.70	63.35	75.80	34.23	37.27	61.43	74.20	31.35	24.81	45.92	57.94	22.78
✓	TIPCB [45]	42.60	64.55	77.38	33.70	39.87	64.01	76.44	33.07	25.41	46.65	57.56	22.50
✓	SSAN [7]	42.36	63.50	76.85	33.54	39.62	63.94	75.10	31.80	26.81	47.55	58.47	23.07
✓	LGUR [6]	44.04	65.63	78.12	35.14	41.67	65.03	77.51	33.32	27.90	48.71	59.96	24.39
×	CCL (w/o FCCL)	47.90	72.05	81.95	36.78	40.75	68.45	79.35	34.07	22.40	47.90	60.25	21.39
×	CCL	**51.05**	**74.70**	**83.00**	**40.31**	**48.40**	**72.90**	**82.65**	**38.92**	**42.15**	**68.10**	**78.15**	**34.47**

TABLE VII
PERFORMANCE COMPARISONS ON CUHK-PEDES UNDER INCOMPLETE TEXT MODE. RESULTS OF EASY, MEDIUM AND HARD SETTINGS ARE REPORTED.

ID	Methods	Easy Setting(50%, 50%)				Medium Setting(30%, 70%)				Hard Setting(10%, 90%)			
		Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10	mAP
✓	CMPM/C [19]	41.22	63.14	74.15	34.88	39.84	62.84	72.25	32.47	26.74	52.33	63.28	24.69
✓	MIA [21]	46.23	67.96	77.42	39.56	43.22	66.06	74.14	36.87	30.23	54.16	64.79	26.96
✓	SCAN [8]	50.13	72.11	79.23	42.55	47.33	69.41	77.67	40.32	32.02	54.25	65.44	28.11
✓	ViTAA [5]	49.14	70.21	79.54	42.23	47.33	70.14	78.57	40.22	31.54	55.21	64.77	27.46
✓	TIPCB [45]	58.66	79.85	87.33	50.74	51.87	74.82	83.04	44.66	21.68	43.74	54.82	19.22
✓	SSAN [7]	53.45	74.68	81.96	45.07	49.25	72.16	80.15	38.88	34.14	57.75	68.33	30.02
✓	LGUR [6]	58.20	79.56	86.81	52.87	52.56	75.21	83.22	48.65	31.45	55.28	66.83	30.06
×	CCL (w/o FCCL)	59.23	80.00	87.54	54.32	50.21	73.08	81.24	44.32	29.14	54.27	66.34	28.45
×	CCL	**64.64**	**84.19**	**90.25**	**58.44**	**62.41**	**82.15**	**88.69**	**56.92**	**55.10**	**76.74**	**84.28**	**50.06**

TABLE VIII
PERFORMANCE COMPARISONS ON ICFG-PEDES UNDER INCOMPLETE TEXT MODE. RESULTS OF EASY, MEDIUM AND HARD SETTINGS ARE REPORTED.

ID	Methods	Easy Setting(50%, 25%, 25%)				Medium Setting(30%, 35%, 35%)				Hard Setting(10%, 45%, 45%)			
		Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10	mAP
✓	CMPM/C [19]	34.73	55.90	65.65	17.59	29.70	51.25	60.80	15.13	15.55	30.09	41.45	8.93
✓	MIA [21]	39.02	60.30	69.20	20.14	35.86	58.40	67.75	18.66	19.70	35.44	46.04	10.57
✓	SCAN [8]	42.48	65.43	72.00	22.07	41.15	63.80	71.71	20.91	22.25	40.33	51.61	11.12
✓	ViTAA [5]	43.85	65.60	73.20	23.30	42.60	64.41	72.33	22.31	22.47	40.60	51.20	11.65
✓	TIPCB [45]	50.24	70.60	77.81	26.51	48.11	67.01	76.49	25.61	28.33	49.47	59.21	14.30
✓	SSAN [7]	46.35	66.86	75.47	24.25	44.80	66.64	74.60	23.86	25.05	42.50	53.09	13.10
✓	LGUR [6]	52.65	70.14	78.48	28.36	48.30	68.65	77.05	26.02	29.81	51.04	59.98	15.00
×	CCL (w/o FCCL)	51.80	71.47	79.12	28.26	44.50	65.81	74.32	23.70	14.10	30.47	39.80	8.03
×	CCL	**53.70**	**73.02**	**80.29**	**29.83**	**51.06**	**70.79**	**78.15**	**27.10**	**41.15**	**62.19**	**71.12**	**22.22**

TABLE IX
PERFORMANCE COMPARISONS ON RSTPREID UNDER INCOMPLETE TEXT MODE. RESULTS OF EASY, MEDIUM AND HARD SETTINGS ARE REPORTED.

ID	Methods	Easy Setting(50%, 25%, 25%)				Medium Setting(30%, 35%, 35%)				Hard Setting(10%, 45%, 45%)			
		Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10	mAP
✓	CMPM/C [19]	33.15	56.75	67.60	27.45	29.38	53.99	65.10	25.85	16.95	38.11	48.68	17.18
✓	MIA [21]	37.55	60.38	71.23	31.23	35.04	58.26	70.73	29.51	21.40	42.93	55.10	20.56
✓	SCAN [8]	40.25	63.31	74.40	32.91	37.65	61.38	73.70	31.41	25.55	47.06	58.10	22.94
✓	ViTAA [5]	40.81	63.46	75.71	34.33	37.21	61.45	74.36	31.40	24.70	46.05	57.87	22.67
✓	TIPCB [45]	42.55	64.55	77.25	33.74	39.78	64.01	76.31	33.14	25.30	46.60	57.63	22.41
✓	SSAN [7]	42.15	63.36	76.68	33.48	39.58	63.60	74.86	31.90	26.80	47.60	58.32	23.22
✓	LGUR [6]	43.95	65.58	78.13	35.10	41.50	64.86	77.40	33.48	27.85	48.69	59.70	24.46
×	CCL (w/o FCCL)	47.68	71.80	81.78	36.66	40.70	68.41	79.36	34.11	22.27	47.60	60.20	21.29
×	CCL	**50.83**	**74.58**	**82.76**	**39.98**	**48.43**	**72.97**	**82.72**	**39.01**	**42.23**	**68.16**	**78.28**	**34.71**

TABLE X
ABLATION STUDY ON EACH COMPONENT OF CCL. PCME AND CCMPM ARE ABLATED UNDER FULL DATA MODE. CSM AND SWG ARE ABLATED UNDER INCOMPLETE DATA MODE.

Exp		Modules								CUHK-PEDES		ICFG-PEDES		RSTPReid	
		PCME	CCMPM	CSM	SWG	Euc	Cos	Cen	Rep	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
Full data	No.0									63.53	57.26	55.15	30.50	47.65	39.44
	No.1	✓								65.38	58.53	56.73	31.26	49.44	40.02
	No.2		✓							65.76	58.61	57.04	31.31	49.83	40.06
	No.3	✓	✓							67.25	60.83	58.33	32.66	51.30	41.10
Incomplete data	No.4	✓	✓							60.05	54.73	50.08	26.97	47.90	36.78
	No.5	✓	✓	✓				✓		61.92	56.52	51.49	28.16	49.15	38.32
	No.6	✓	✓	✓					✓	59.53	54.22	49.67	26.61	47.55	36.31
	No.7	✓	✓		✓	✓				60.98	56.10	50.78	27.86	48.52	37.96
	No.8	✓	✓		✓		✓			61.40	55.83	51.10	27.72	48.80	37.73
	No.9	✓	✓	✓	✓					64.76	58.81	53.64	29.70	51.05	40.31

text mode. Likewise, we compared it with the open-sourced methods, and the results are shown in Table VII, Table VIII and Table IX. Similarly, the performance of comparative methods and 'CCL (w/o FCCL)' decrease significantly when only a small amount of text descriptions are used, while CCL still achieves competitive performance. On CUHK-PEDES, with only 10% of the text descriptions used, CCL outperforms CMPM/C and MIA trained with full data, with 55.10% achieved on Rank-1, 76.74% achieved on Rank-5, 84.28% achieved on Rank-10 and 50.06% achieved on mAP. On RSTPReid, with only 10% of the text descriptions used, CCL outperforms IMG-Net, AMEN, DSSL and SUM trained with full data, with 42.23% achieved on Rank-1, 68.16% achieved on Rank-5, 78.28% achieved on Rank-10 and 34.71% achieved on mAP. All these results demonstrate the significant effectiveness of CCL in reducing the use of annotated texts.

Discussion. Comparing the results of Table IV, Table V, Table VI and Table VII, Table VIII, Table IX, we can observe that the performance of CCL is similar. This is because while addressing the problem of incomplete data and reducing annotation costs using a small amount of descriptions are distinct goals, they are essentially similar in nature under the proposed framework. CCL completes all incomplete data and involves it in the training process. The case of using a small amount of descriptions can be seen as a special case of incomplete data, where the missing data are all text descriptions. Therefore, CCL provides a general solution for both scenarios, making it more convenient and flexible for practical applications.

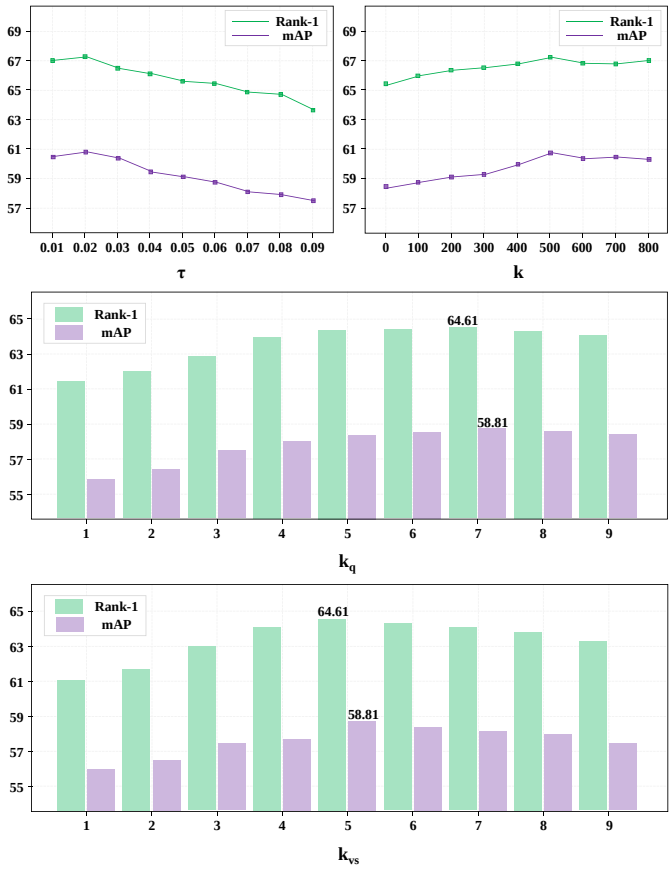

Fig. 7. Effect of τ, k, k_q, and k_{vs} on CUHK-PEDES. k represents the number of cross-modal shared prototypes in PCME. τ is the temperature coefficient of the CCMPM loss. k_q represents the number of mutual neighbors obtained in CSM, and k_{vs} represents the number of semantically most similar features used to generate missing features in SWG.

C. Ablation Studies

To demonstrate the effectiveness of each module and setting in the proposed framework, we conduct detailed ablation experiments on CCL, as shown in Table X. In the table, the No.0 experiment represents our baseline, where we train using InfoNCE loss. Experiments No.1-No.3 are tested under the full data mode, while experiments No. 4-No.9 are tested under the incomplete data mode. 'Euc' denotes Euclidean distance, 'Cos' represents cosine similarity, 'Cen' denotes the method of taking the centroid of obtained features as a replacement, and 'Rep' means the method of directly using the most similar feature as a replacement.

The effectiveness of PCME. Comparing experiments No.0 and No.1, as well as experiments No.2 and No.3, demonstrates the effectiveness of PCME in semantic alignment and reducing modality discrepancy. In iTIReID, explicit local semantic alignment in previous methods is difficult to achieve due to the lack of identity labels and some data. PCME performs implicit local semantic alignment and reduces modality differences, mapping features of different modalities to a shared latent space, making it easier to measure distances between features

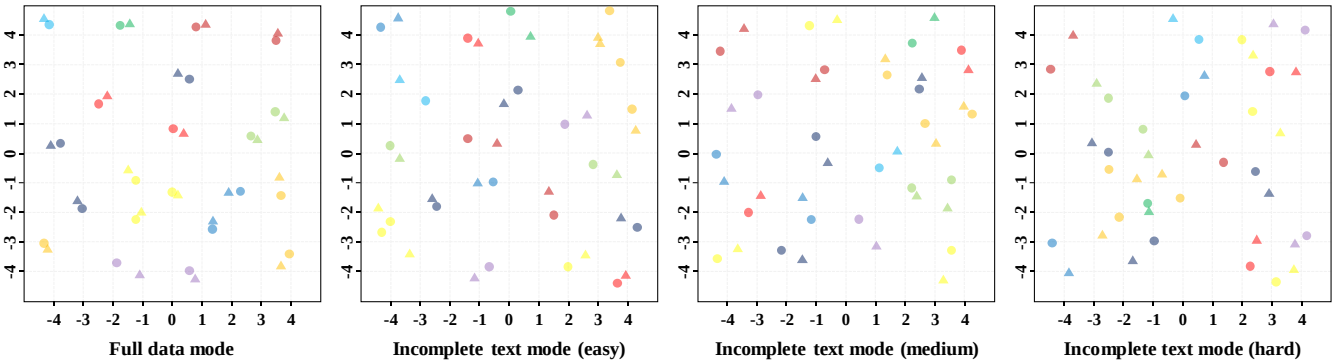

Fig. 8. t-SNE of the the generated features under different difficulty settings of incomplete data mode. To facilitate comparison, t-SNE of the features under the full data mode is given as well. Triangles represent features of image modality and circles represent features of text modality. Features of the same identity are marked with the same color.

in subsequent steps. Specifically, comparing experiments No.2 and No.3, PCME promotes Rank-1 by 1.49%, Rank-5 by 1.41%, Rank-10 by 1.35% and mAP by 2.22%.

The effectiveness of the CCMPM loss. Comparing experiments No.0 and No.2 as well as No.1 and No.3 demonstrates the effectiveness of CCMPM loss in weakly supervised contrastive training. In iTIReID, the identity-based optimization commonly used in the existing methods is no longer applicable due to the lack of identity labels. The CCMPM loss combines the advantages of InfoNCE loss and CMPM loss, and can effectively utilize the correspondence between images and text descriptions for contrastive learning training, maximizing the similarity between positive samples and minimizing the similarity between negative samples, forming a reasonable feature distribution where only pairwise samples have absolute high similarity. Compared with InfoNCE loss, CCMPM increases Rank-1 by 2.23%, Rank-5 by 1.58%, Rank-10 by 0.34% and mAP by 1.35% on the baseline.

The effectiveness of CSM. The effectiveness of CSM in measuring cross-modal semantic similarity is validated through comparative experiments No. 7 and No. 9, as well as experiments No. 8 and No. 9. Due to the existence of modality discrepancy, directly adopting Euclidean distance or cosine similarity to measure cross-modal semantic similarity is not reliable. CSM avoids modality discrepancy by leveraging intra-modality similarity when measuring cross-modal similarity, in order to select the most semantically similar features. Compared to cosine similarity, CSM improves Rank-1 by 3.36%, Rank-5 by 2.69%, Rank-10 by 0.51% and mAP by 2.98% when measuring cross-modal similarity.

The effectiveness of SWG. The effectiveness of SWG in generating features by semantic relevance weight is demonstrated through comparative experiments No. 5 and No. 9, as well as No. 6 and No. 9. The methods for cross-modal retrieval tends to use the centroid of relevant features as the semantic center or cluster center, but this approach is not suitable for iTIReID. Instead, SWG generates features based on semantic similarity weight, where features with higher similarity contribute more to the generated features. Compared to the approach of using the centroid as a replacement for missing features, SWG improves Rank-1 by 2.84%, Rank-5 by 2.95%, Rank-10 by 0.48% and mAP by 2.29%.

Hyperparameter analysis. In the proposed framework, we adopt four hyperparameters, τ, k, k_q, and k_{vs}. We varied τ, k, k_q, and k_{vs} within certain ranges and recorded the changes in the Rank-K and mAP metric to analyze their optimal values on CUHK-PEDES. When analyzing one hyperparameter, the other hyperparameters remained constant. The hyperparameter analysis chart is shown in Fig. 7. k represents the number of cross-modal shared prototypes in PCME. It can be observed that when $k = 500$, the model performance tends to saturate, meaning that the number of prototypes at this point is sufficient to learn strong cross-modal shared features. τ is the temperature coefficient of the CCMPM loss. It can be observed that the optimal value for τ is 0.02. As τ increases, the model performance decreases more. k_q represents the number of mutual neighbors obtained in CSM, and k_{vs} represents the number of semantically most similar features used to generate missing features in SWG. It can be observed that when k_q is small, CSM cannot fully function, and when k_q is large, CSM exhibits weak discrimination. Similarly, when k_{vs} is small, the features generated by SWG are not representative, and when k_{vs} is large, the quality of the generated features decreases. Therefore, the experiments show that in CCL, the optimal values for τ, k, k_q, and k_{vs} are 0.02, 500, 7, and 5, respectively.

D. Feature Visualization

In order to better demonstrate the effectiveness and rationality of the generated features, we select some of the generated features under different difficulty settings of incomplete data mode on CUHK-PEDES and plot their t-SNE. To facilitate comparison, we also plot the t-SNE of the features under the full data mode, as shown in Fig. 8. It is worth noting that half of the features are generated features. We can observe that under the full data mode, the features form very reasonable feature embeddings under the optimization of the CCMPM loss, and the features of the corresponding image and text modes cluster together, with compact clusters and clear boundaries. We also observe the same phenomenon under the incomplete data mode, especially in simple and medium difficulty settings, which confirms the effectiveness

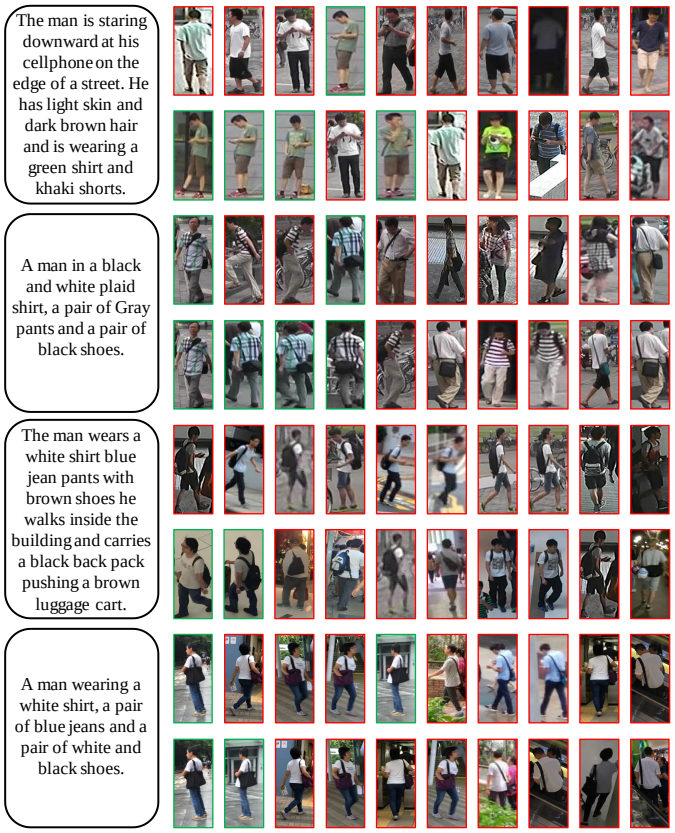

Fig. 9. Comparison of top-10 retrieved results on CUHK-PEDES (incomplete data mode, hard setting) between CCL w/o FCCL (the first row) and CCL (the second row). The images corresponding to matched and mismatched images are marked with green and red rectangles, respectively.

of the generated features in CCL. As the difficulty level increases and more incomplete data is added, the clusters formed become slightly looser, but the generated features still have a correct and strong correspondence with the original modal features, which verifies the effectiveness of CCL in dealing with extreme situations.

E. Qualitative Results

Fig. 9 presents a comparison of the top-10 retrieval results for CCL w/o FCCL (the first row) and CCL (the second row) under hard setting on the CUHK-PEDES dataset. In Fig. 9, We can observe that CCL effectively utilizes incomplete data to achieve more accurate retrieval results. This is attributed to the rationality of the feature completion strategy and the effectiveness of the CSM and SWG. Furthermore, it can be observed from the figure that CCL is able to distinguish well among difficult samples with similar appearances, primarily due to the PCME's ability to align local semantic implicitly. The qualitative results intuitively demonstrate the robustness of CCL in real-world scenarios.

V. Conclusion

In this paper, we propose for the first time incomplete Text-Image person re-identification (iTIReID), which comprises a small amount of complete pairwise data and a large amount of incomplete data, where all identity labels are unavailable. We introduce a novel Contrastive Completing Learning (CCL) framework for iTIReID, consisting of two stages: Pure Contrastive Learning (PCL) and Feature Completion Contrastive Learning (FCCL). In PCL, only complete pairwise data is utilized for training, which serves as a preliminary improvement of the model's capacity and prepares for the upcoming feature completion stage. In FCCL, available features are used to complete missing modality features and facilitate effective training with incomplete data. During this process, Cross-modal Semantic Measure (CSM) is proposed to leverage intra-modality similarity to measure cross-modal similarity and filter out features with the highest semantic similarity, thereby circumventing modality discrepancy. Semantic-Weighted Generation (SWG) is proposed to generate features based on the semantic similarity weight of the similar features. To fully leverage pairwise data for label-free training, we introduce the contrastive CMPM (CCMPM) loss for contrastive learning to achieve weakly supervised training. Experimental results verify the effectiveness of our proposed methods and demonstrate competitive performance compared to fully supervised methods using complete data.

References

[1] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. Hoi, "Deep learning for person re-identification: A survey and outlook," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 6, pp. 2872–2893, 2021.

[2] J. Zhu, H. Zeng, S. Liao, Z. Lei, C. Cai, and L. Zheng, "Deep hybrid similarity learning for person re-identification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 11, pp. 3183–3193, 2017.

[3] S. Li, T. Xiao, H. Li, B. Zhou, D. Yue, and X. Wang, "Person search with natural language description," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1970–1979.

[4] L. Gao, K. Niu, B. Jiao, P. Wang, and 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*, 2023.

[5] Z. Wang, Z. Fang, J. Wang, and Y. Yang, "Vitaa: Visual-textual attributes alignment in person search by natural language supplementary material."

[6] Z. Shao, X. Zhang, M. Fang, Z. Lin, J. Wang, and 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.

[7] Z. Ding, C. Ding, Z. Shao, and D. Tao, "Semantically self-aligned network for text-to-image part-aware person re-identification," *arXiv preprint arXiv:2107.12666*, 2021.

[8] K.-H. Lee, X. Chen, G. Hua, H. Hu, and X. He, "Stacked cross attention for image-text matching," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 201–216.

[9] Z. Zheng, L. Zheng, and Y. Yang, "Pedestrian alignment network for large-scale person re-identification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 10, pp. 3037–3045, 2018.

[10] P. Li, P. Pan, P. Liu, M. Xu, and Y. Yang, "Hierarchical temporal modeling with mutual distance matching for video based person re-identification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 2, pp. 503–511, 2020.

[11] L. Qi, L. Wang, J. Huo, Y. Shi, and Y. Gao, "Progressive cross-camera soft-label learning for semi-supervised person re-identification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 9, pp. 2815–2829, 2020.

[12] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," *arXiv preprint arXiv:1703.07737*, 2017.

[13] W. Chen, X. Chen, J. Zhang, and 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.

[14] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and 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.

[15] G. Wang, Y. Yuan, X. Chen, J. Li, and X. Zhou, "Learning discriminative features with multiple granularities for person re-identification," in *Proceedings of the 26th ACM international conference on Multimedia*, 2018, pp. 274–282.

[16] F. Faghri, D. J. Fleet, J. R. Kiros, and S. Fidler, "Vse++: Improving visual-semantic embeddings with hard negatives," *arXiv preprint arXiv:1707.05612*, 2017.

[17] N. Sarafianos, X. Xu, and 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.

[18] Z. Zheng, L. Zheng, M. Garrett, Y. Yang, M. Xu, and Y.-D. Shen, "Dual-path convolutional image-text embeddings with instance loss," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 16, no. 2, pp. 1–23, 2020.

[19] Y. Zhang and H. Lu, "Deep cross-modal projection learning for image-text matching," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 686–701.

[20] T. Chen, C. Xu, and 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.

[21] K. Niu, Y. Huang, W. Ouyang, and L. Wang, "Improving description-based person re-identification by multi-granularity image-text alignments," *IEEE Transactions on Image Processing*, vol. 29, pp. 5542–5556, 2020.

[22] D. Jiang and M. Ye, "Cross-modal implicit relation reasoning and aligning for text-to-image person retrieval," *arXiv preprint arXiv:2303.12501*, 2023.

[23] Y. Patel, L. Gomez, M. Rusiñol, D. Karatzas, and C. Jawahar, "Self-supervised visual representations for cross-modal retrieval," in *Proceedings of the 2019 on International Conference on Multimedia Retrieval*, 2019, pp. 182–186.

[24] R. Gomez, L. Gomez, J. Gibert, and D. Karatzas, "Self-supervised learning from web data for multimodal retrieval," in *Multimodal Scene Understanding*. Elsevier, 2019, pp. 279–306.

[25] S. Zhao, C. Gao, Y. Shao, W.-S. Zheng, and N. Sang, "Weakly supervised text-based person re-identification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 11 395–11 404.

[26] L. Wu, Y. Wang, and L. Shao, "Cycle-consistent deep generative hashing for cross-modal retrieval," *IEEE Transactions on Image Processing*, vol. 28, no. 4, pp. 1602–1612, 2018.

[27] J. Guo and W. Zhu, "Collective affinity learning for partial cross-modal hashing," *IEEE Transactions on Image Processing*, vol. 29, pp. 1344–1355, 2019.

[28] Z. Zeng, S. Wang, N. Xu, and W. Mao, "Pan: Prototype-based adaptive network for robust cross-modal retrieval," in *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021, pp. 1125–1134.

[29] J. Liu, Y. Sun, F. Zhu, H. Pei, Y. Yang, and W. Li, "Learning memory-augmented unidirectional metrics for cross-modality person re-identification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 19 366–19 375.

[30] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[31] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[32] 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.

[33] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *International Conference on Machine Learning*. PMLR, 2022, pp. 12 888–12 900.

[34] Y. Lu, Y. Wu, B. Liu, T. Zhang, B. Li, Q. Chu, and N. Yu, "Cross-modality person re-identification with shared-specific feature transfer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 13 379–13 389.

[35] G. Wang, T. Zhang, J. Cheng, S. Liu, Y. Yang, and Z. Hou, "Rgb-infrared cross-modality person re-identification via joint pixel and feature alignment," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 3623–3632.

[36] Z. Wang, Z. Wang, Y. Zheng, Y.-Y. Chuang, and S. Satoh, "Learning to reduce dual-level discrepancy for infrared-visible person re-identification," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 618–626.

[37] Z. Zhong, L. Zheng, D. Cao, and S. Li, "Re-ranking person re-identification with k-reciprocal encoding," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1318–1327.

[38] A. Zhu, Z. Wang, Y. Li, X. Wan, J. Jin, T. Wang, F. Hu, and G. Hua, "Dssl: deep surroundings-person separation learning for text-based person retrieval," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 209–217.

[39] Z. Wang, A. Zhu, Z. Zheng, J. Jin, Z. Xue, and G. Hua, "Img-net: inner-cross-modal attentional multigranular network for description-based person re-identification," *Journal of Electronic Imaging*, vol. 29, no. 4, pp. 043 028–043 028, 2020.

[40] Z. Wang, A. Zhu, J. Xue, D. Jiang, C. Liu, Y. Li, and F. Hu, "Sum: Serialized updating and matching for text-based person retrieval," *Knowledge-Based Systems*, vol. 248, p. 108891, 2022.

[41] S. Yan, H. Tang, L. Zhang, and J. Tang, "Image-specific information suppression and implicit local alignment for text-based person search," *arXiv preprint arXiv:2208.14365*, 2022.

[42] X. Shu, W. Wen, H. Wu, K. Chen, Y. Song, R. Qiao, B. Ren, and X. Wang, "See finer, see more: Implicit modality alignment for text-based person retrieval," in *Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part V*. Springer, 2023, pp. 624–641.

[43] Z. Wang, A. Zhu, J. Xue, X. Wan, C. Liu, T. Wang, and 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.

[44] X. Han, S. He, L. Zhang, and T. Xiang, "Text-based person search with limited data," *arXiv preprint arXiv:2110.10807*, 2021.

[45] Y. Chen, G. Zhang, Y. Lu, Z. Wang, and Y. Zheng, "Tipcb: A simple but effective part-based convolutional baseline for text-based person search," *Neurocomputing*, vol. 494, pp. 171–181, 2022.

[46] Z. Wang, A. Zhu, J. Xue, X. Wan, C. Liu, T. Wang, and 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.

[47] A. Farooq, M. Awais, J. Kittler, and S. S. Khalid, "Axm-net: Implicit cross-modal feature alignment for person re-identification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 4, 2022, pp. 4477–4485.

[48] Z. Wang, J. Xue, A. Zhu, Y. Li, M. Zhang, and C. Zhong, "Amen: Adversarial multi-space embedding network for text-based person re-identification," in *Pattern Recognition and Computer Vision: 4th Chinese Conference, PRCV 2021, Beijing, China, October 29–November 1, 2021, Proceedings, Part II 4*. Springer, 2021, pp. 462–473.

[49] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

Guodong Du received his B.S. degree from Nanjing University of Aeronautics and Astronautics, NanJing, China, in 2022. He is currently pursuing Ph.D. degree in Nanjing University of Aeronautics and Astronautics, NanJing, China, since 2022. His research interests include Computer vision and Multi-media.

Tiantian Gong is currently pursuing the Ph.D degree at School of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing, China. Her research interests include computer vision and multimedia analysis.

Liyan Zhang received the PhD degree in computer science from the University of California, Irvine, in 2014. She is currently a professor at the School of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, China. Her research interests include multimedia analysis and computer vision. She has received the Best Paper Award in ICMR 2013 and the Best Student Paper Award in MMM 2016.