Text-Based Occluded Person Re-identification via Multi-Granularity Contrastive Consistency Learning

Xinyi Wu[1], Wentao Ma[2]*, Dan Guo[3], Tongqing Zhou[1]*, Shan Zhao[3], Zhiping Cai[1]*

[1]College of Computer, National University of Defense Technology, Changsha, China
[2]School of Information and Artificial Intelligence, Anhui Agricultural University, Hefei, China
[3]School of Computer Science and Information Engineering, HeFei University of Technology, Hefei, China
{wuxinyi17, zhoutongqing, zpcai}@nudt.edu.cn, wtma@ahau.edu.cn, {guodan, zhaoshan}@hfut.edu.cn

Abstract

Text-based Person Re-identification (T-ReID), which aims at retrieving a specific pedestrian image from a collection of images via text-based information, has received significant attention. However, previous research has overlooked a challenging yet practical form of T-ReID: dealing with image galleries mixed with occluded and inconsistent personal visuals, instead of ideal visuals with a full-body and clear view. Its major challenges lay in the insufficiency of benchmark datasets and the enlarged semantic gap incurred by arbitrary occlusions and modality gap between text description and visual representation of the target person. To alleviate these issues, we first design an Occlusion Generator (OGor) for the automatic generation of artificial occluded images from generic surveillance images. Then, a fine-granularity token selection mechanism is proposed to minimize the negative impact of occlusion for robust feature learning, and a novel multi-granularity contrastive consistency alignment framework is designed to leverage intra-/inter-granularity of visual-text representations for semantic alignment of occluded visuals and query texts. Experimental results demonstrate that our method exhibits superior performance. We believe this work could inspire the community to investigate more dedicated designs for implementing T-ReID in real-world scenarios. The source code is available at https://github.com/littlexinyi/MGCC.

Introduction

Person Re-identification (ReID) is a fundamental yet challenging task in computer vision (CV), playing a paramount role in plenty of applications such as intelligent video surveillance, urban security, and smart retailing (Ye et al. 2021; Zeng et al. 2022).

However, existing person ReID methods (Yao et al. 2019; Ding et al. 2020; Wang et al. 2021b) usually use images of a specific person as the probes, which have limitations in real-world urgent scenarios. For instance, when police officers try to locate criminals (suspects) or lost children within a shopping mall, they typically lack any photographic references of the individuals. Fortunately, they can take verbal descriptions of the target person from witnesses and cross-check them with surveillance videos. To release manually checking, a novel and feasible paradigm for person ReID, text-based person ReID (T-ReID) (Li et al. 2017; Zhu et al. 2021; Wang et al. 2020b; Ma et al. 2023), has been proposed recently. As shown in Figure 1(a), the cross-modal correlations are established between verbal descriptions and widely monitored images, to remedy the absence of visual cues in real-world scenarios.

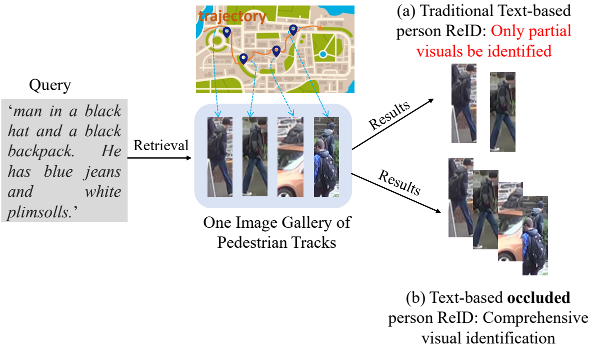

Query
'man in a black hat and a black backpack. He has blue jeans and white plimsolls.'
Retrieval
Trajectory
One Image Gallery of Pedestrian Tracks
Results
Results
(a) Traditional Text-based person ReID: Only partial visuals be identified
(b) Text-based **occluded** person ReID: Comprehensive visual identification

Figure 1: Illustration of person ReID: (a) Text-based person ReID and (b) Text-based occluded person ReID.

Generally, T-ReID needs to process visual and textual modalities, which is aimed at learning a common semantic representation space between visual and textual modalities, to better align image and text. For this purpose, recent works firstly employ different models to extract the feature representations of different modalities from local feature level (Chen et al. 2022), global feature level (Chen et al. 2021b), and multi-granularity feature level (Farooq et al. 2021; Wang et al. 2022a). Then they focus on exploring image-text pairs for semantic alignment in common representation space.

Unfortunately, we observe that the model training process employed by these methods on existing public benchmark datasets (Li et al. 2017; Ding et al. 2021; Zhu et al. 2021) is based on a strong assumption: *the information on image modality is holistic (without occlusions), namely, the model can extract effective feature representations of images*. This is definitely unrealistic, wherein the person ReID techniques currently face a fundamental and long-standing

*Corresponding Authors

challenge: *Occlusions*. Although occluded person ReID has been widely investigated (Zhuo et al. 2018; He and Liu 2020; Wang et al. 2020a; Zheng et al. 2021), they are merely deployed in the visual-visual retrieval scenarios, hard to be adapted to *complex cross-modal person ReID*. Therefore, how to effectively tackle the complex occlusion issues and realize adaption to real-world scenarios has become one of the key challenges in T-ReID.

In light of the above analysis, in this paper, we attempt to explore *a novel and advanced case of T-ReID*, *text-based occluded person ReID (TO-ReID)*, as shown in Figure 1(b). However, the implementation process faces the following three challenges:

1. **Absence of benchmark datasets.** Most existing datasets of T-ReID implicitly assume that the appearance of full-body for a person is readily available, while ignoring person images with occlusions, hindering models from acquiring robust visual features that are suitable for TO-ReID task.
2. **Semantic gap from occlusion.** Occlusion causes key feature loss and redundant feature interference, which will introduce huge semantic gap for image feature learning.
3. **Modality gap for image-text pairs.** Due to the inconsistent representation of images and text, their data resides in different distribution spaces, making it difficult to directly measure the similarity between image-text pairs.

To relieve the above challenges, we first design an Occlusion Generator (termed OGor) to support the benchmark dataset construction for TO-ReID, which automatically generates artificial occluded person images by occlusion sample augmentation strategy (Challenge 1). Then, a fine-granularity token selection mechanism is proposed, to eliminate redundant noisy tokens stemming from occlusions and meaningless auxiliary words. This mechanism serves two pivotal objectives: bridging the semantic gap induced by occlusions and realizing the trade-off between performance and training efficiency (Challenge 2).

Finally, a multi-granularity contrastive consistency learning model is also invented, enabling better alignment for text-image pairs (Challenge 3). The main contributions of this paper can be summarized as follows:

- We invent an OGor and reconstruct the original T-ReID dataset to simulate real-world surveillance occlusion scenarios. Additionally, the OGor is versatile and can be extended to other general datasets, enhancing the exploration of novel scenarios.
- We propose a **M**ulti-**G**ranularity **C**ontrastive **C**onsistency model, dubbed **MGCC**, incorporating a token selection mechanism. This model not only bridges the semantic gap arising from occlusion but also narrows the modality gap between images and texts.
- Extensive experiments demonstrate the effectiveness of our proposal on three TO-ReID datasets, *i.e.*, Occluded-CUHK-PEDES (62.44 on R@1), Occluded-ICFG-PEDES (59.28 on R@1), and Occluded-RSTPReid (49.85 on R@1).

Related Work

Occlusion Person ReID

Person ReID has been widely investigated; however, occlusions degrade the robustness of feature representation, leading to a decay in performance. Several works (Zhuo et al. 2018; Zheng et al. 2021; Zhou et al. 2023) attempt to tackle the complex visual occlusions, which can be roughly categorized into two groups: part-to-part matching (Sun et al. 2018; He et al. 2018, 2019; He and Liu 2020; Zhu et al. 2020) and matching by external tools assistance (Miao et al. 2019; Gao et al. 2020; Wang et al. 2020a; Zheng et al. 2021). For the former, it realizes matching by measuring the similarity between local features (*e.g.*, body parts).

Zhu et al. (2020) propose an identity-guided human semantic parsing approach, ISP, to generate the pseudo-labels of human body parts at pixel-level. For the latter, techniques such as pose estimation and human parsing are used to realize the alignment. Miao et al. (2019) propose a Pose-Guided Feature Alignment model, called PGFA, which leverages the human pose key points for matching. To achieve high accuracy while preserving low inference complexity, Zheng et al. (2021) propose a PGFL-KD model, which can distill human pose semantic knowledge into a local feature extractor to discard the dependency on a pose estimator.

While these techniques have shown effectiveness in improving the performance of occluded person ReID, their application is still confined to image-based person ReID methods. Notably, in real-world scenarios that involve text-based person ReID, the challenge of occlusion has not been specifically addressed.

Text-based Person ReID

Considering that visual cues are not always available in real-world scenarios, we focus on the realistic problem of T-ReID. As given by the pioneering work (Li et al. 2017) that introduces the T-ReID task and releases a benchmark dataset named CUHK-PEDES, the main challenge for this task is how to efficiently align image and text features into a joint embedding space for fast retrieval. Solutions can be categorized into two groups: single-granularity feature alignment paradigm and multi-granularity feature alignment paradigm.

In terms of the single-granularity, major improvements are shown by (Chen et al. 2022; Zhang and Lu 2018; Wang et al. 2019; Liu et al. 2019; Ge, Gao, and Liu 2019; Zheng et al. 2020b; Chen et al. 2021b) where researchers start exploiting local alignment and global alignment. (Chen, Xu, and Luo 2018) propose a patch-word level matching for T-ReID, to realize local feature alignment. To reduce the additional modules and complex evaluation strategies, (Chen et al. 2022) design a simple but effective framework, TIPCB, for T-ReID via learning visual and textual local representations. While other works employ global feature representation to realize global matching.

For the multi-granularity (Jing et al. 2020; Wang et al. 2020b; Zheng et al. 2020a; Zhu et al. 2021; Wang et al. 2021a; Ding et al. 2021; Shao et al. 2022; Wang et al. 2022b; Jiang and Ye 2023), this paradigm brings decent improvements as compared to using single-granularity. Among them, Niu et al. (2020) propose an end-to-end multi-granularity image-text alignment, MIA, to extract fine-grained features of three different granularities hierarchically and then adopt a cross-modal attention mechanism to determine affinities between visual and textual components.

Wang et al. (2020b) present a ViTAA model, which learns to disentangle the feature space of a person into subspaces corresponding to attributes, to address the T-ReID task from the perspective of attribute-specific alignment learning. Owing to the significant modality gap and the large intra-class variance in texts, a model called SSAN is designed by (Ding et al. 2021), which can automatically extract semantically aligned features from the visual and textual modalities. Recently, a model called IRRA (Jiang and Ye 2023) is proposed which focuses on cross-modal implicit relation reasoning and aligning for T-ReID tasks.

Vision-Language Pre-Training Models

The pre-training and fine-tuning paradigm has achieved great success, which drives the development of CV (Dosovitskiy et al. 2020) and natural language processing (NLP) (Brown et al. 2020). Many efforts (Yan et al. 2022; Radford et al. 2021; Ma et al. 2022; Yao et al. 2021; Cao et al. 2022; Fang et al. 2021; Shu et al. 2022) have attempted to extend the pre-training model to the multimodal field. It is gratifying that visual language pre-training (VLP) has attracted growing attention. Among them, CLIP (Radford et al. 2021) has gained surging popularity.

As a leading pre-training model, different from the traditional single-modality supervised pre-training model, CLIP leverages natural text descriptions to supervise the learning. Since the great advantage of CLIP, a lot of follow-ups (Luo et al. 2022; Fang et al. 2021; Ma et al. 2022; Zhao et al. 2022; Shu et al. 2022; Han et al. 2021; Yan et al. 2022) have also begun to transfer the knowledge of CLIP to visual-textual retrieval tasks and obtained new state-of-the-art (SOTA) results. Of course, as a specific application for image-text cross-modal retrieval, T-ReID can also benefit from CLIP. Accordingly, we explore leveraging CLIP for the TO-ReID task. To the best of our knowledge, this is the first attempt to harness CLIP to settle the occlusions of pedestrians in T-ReID.

Dataset Design

To facilitate research on TO-ReID, we design an occlusion generator, termed OGor, which is applied to the existing three datasets of T-ReID (Li et al. 2017; Ding et al. 2021; Zhu et al. 2021) to construct new occluded datasets for TO-ReID.

Different from existing Random Erase (Zhong et al. 2020) and Random Cropping (Chen et al. 2021a) method with weak generation ability while facing diversified occlusions, our OGor adopts an occlusion sample augmentation strategy with realistic occlusion scenario simulation, which mainly contains the following two steps:

Occlusion Instance Library Collection

We establish an occlusion instance library (OIL) by detecting 60 common occlusion samples (4 samples per class, a total of 15 classes) in common outdoor scenes. Specifically, we utilize Mask R-CNN (He et al. 2017) to identify the occlusion instance bounding box and subsequently erase the extraneous background pixels to produce the occlusion instance masks. *15 samples of occluded instance images (select one sample as a representative for each class) from OIL will be shown in Appendix (§A).*

Up Instance Set O_U	Middle Instance Set O_M	Bottom Instance Set O_B
umbrella, kite.	bag, suitcase, post.	car, bike, stone, motorbike, bench, road sign, chair, card, pedestrian, fire hydrant.

Table 1: The list of occlusion instances in OIL.

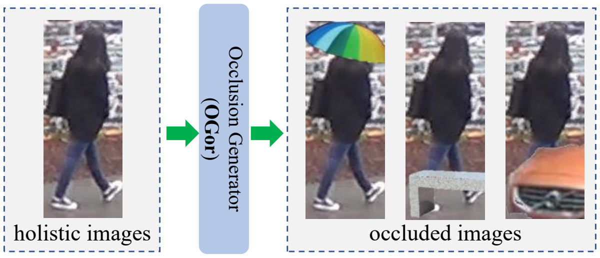


Figure 2: Sample occluded images are generated via OGor.

Occlusion Generation Process

Based on empirical observations, certain common occlusions tend to have position priors in detected person image (*e.g.*, cars, bikes, and pedestrians are generally in the lower half of the image and are unlikely to occur elsewhere).

As a result, according to the classes of occlusion, we divide the OIL into three subsets (as illustrated in Table 1). For the Bottom Instance Set, we align the bottom edge and place them randomly in the horizontal directions. For the Up Instance Set, we align the up edge and randomly place occlusions in horizontal directions. Instances in the Middle Instance Set are not limited to any specific locations, but rather encompass generalized intermediate regions.

To generate occluded images, we randomly select 30% of the whole train, val, and test images within the same ID but different views from the T-ReID dataset. For each selected image, we paste the occlusion instances onto the corresponding regions of the image according to the different position of occlusions in Table 1. *The detailed occlusion generation algorithm is described in Appendix (§B).*

Methodology

In this section, we elaborate the proposed MGCC model from the following five parts.

Multi-Granularity Feature Representations

Image feature representation. For a given image $I_k \in \mathbf{I}$, we first divide the image into n patches and introduce a vi-

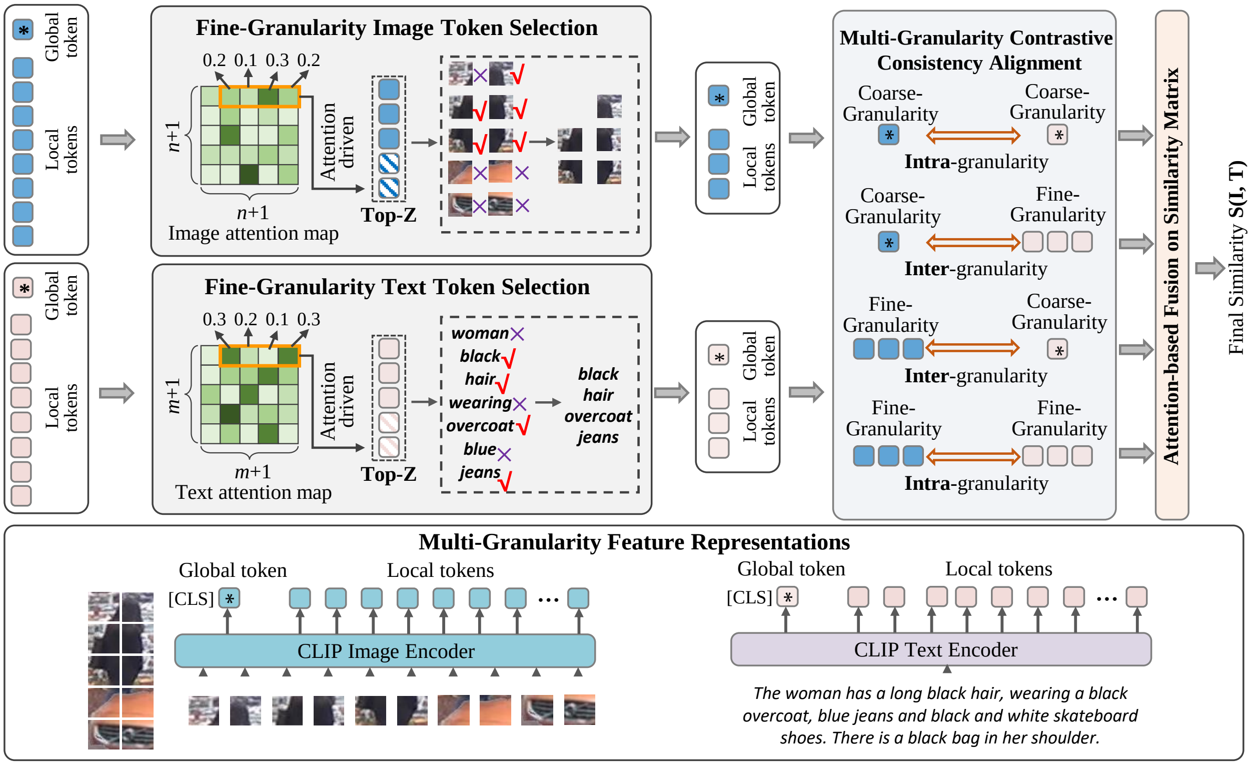


Figure 3: An overview of our MGCC model. First, the text encoder and image encoder extract multi-granularity feature representations for image and text, respectively. Then, a fine-granularity token selection mechanism is applied for filtering the Top-Z informative tokens. Building upon these feature representations, we leverage two types of granularity features to engage in contrastive learning within a common semantic space.

sual tokenization operation to generate a discrete token sequence. A learnable [CLS] token is attached to the beginning of the sequence as an image-level representation. Then, we finetune the standard CLIP with 12 layers as an image encoder to embed the discrete tokens into image-level and patch-level feature representations. To be specific, the [CLS] token is learned as coarse-granularity global feature representations $I_k^g \in \mathbb{R}^{dim}$, and the other learned tokens are regarded as fine-granularity local feature representations $P_k^l = \left[p_{(k,1)}, p_{(k,2)}, p_{(k,3)}, \dots, p_{(k,n)}\right] \in \mathbb{R}^{n \times dim}$.

Text feature representation. Given a text sentence $T_k \in \mathbf{T}$, the raw text is firstly tokenized by CLIP Tokenizer, then the textual sequence is padded with a [CLS] token at the beginning and fed into the text encoder. Similar to the image encoder, the text encoder is also initialized by the public CLIP checkpoints to generate textual feature representations. As a result, we can get coarse-granularity global feature representations $T_k^g \in \mathbb{R}^{dim}$ and fine-granularity local feature representations $\mathrm{W}_k^l = \left[w_{(k,1)}, w_{(k,2)}, w_{(k,3)}, \dots, w_{(k,m)}\right] \in \mathbb{R}^{m \times dim}$, which are respectively the embedding representation of the [CLS] token and corresponding word tokens, where m is the length of word tokens.

Fine-Granularity Token Selection Mechanism

Considering the feature redundancy caused by occlusions on the person image and punctuation marks or meaningless words on the whole sentence, we further design a token selection mechanism for image and text feature representations at fine-granularity levels, which aims at pruning redundant tokens via information importance ranking.

Specifically, we utilize the Transformer block of encoders, where the self-attention of the last block can generate an attention map $M \in \mathbb{R}^{(1+n/m)\times(1+n/m)}$, which reflects the correlation among tokens. We select the first row of the attention map as importance scores between the [CLS] token and all fine-granularity tokens, which is defined as $M' = M[0, 1:] \in \mathbb{R}^{n/m}$. A larger $M'[i]$ means a greater contribution from the i-th fine-granularity feature representation, thus selecting the Top-Z informative tokens from the raw tokens to participate in training and inference while masking the other unimportant tokens, which can realize the trade-off between competitive complexity and performance at the same time.

The token selection mechanism is applied separately to images and text. For a given image I_k with n patches, top Z_i informative tokens are selected from fine-granularity feature representation P_k^l. For a given text T_k with m words, top Z_t informative tokens are selected from fine-granularity feature representation W^l_k. $Z_i = \rho_i \times n$ and $Z_t = \rho_t \times m$, where ρ_i and ρ_t represent the token selection ratio for images and texts, respectively. *Detailed ablation studies of ρ_i and ρ_t are presented in Appendix (§F).*

Multi-Granularity Contrastive Consistency Alignment

To learn high-quality representation alignment, we perform our proposed model at intra-granularity parallel contrast level and inter-granularity cross contrast level. For clarity and simplicity, we have omitted the indexes of all feature representations.

Suppose that there is a batch of training pairs, $B = \{(I_k, T_k)\}_{k=1}^{N}$, and we have extracted feature representations of coarse-to-fine granularity for them.

Intra-granularity parallel contrast alignment. We calculate the similarity of image-text on fine-granularity and coarse-granularity by matrix multiplication, respectively. Consequently, in terms of fine-granularity similarity, we can formulate it as:

$$S_{P-W} = PW^{\top}, \tag{1}$$

where $S_{P-W} \in \mathbb{R}^{n \times m}$ is the similarity score on fine-granularity.

And given image feature at coarse-granularity $I \in \mathbb{R}^{dim}$ and text feature at coarse-granularity $T \in \mathbb{R}^{dim}$, then the similarity matrix of coarse-granularity can be obtained using the matrix multiplication:

$$S_{I-T} = I^{\top}T, \tag{2}$$

where $S_{I-T} \in \mathbb{R}^1$ is the similarity score on coarse-granularity.

Inter-granularity cross contrast alignment. Given the image feature at fine-granularity $P \in \mathbb{R}^{n \times dim}$ and text feature at coarse-granularity $T \in \mathbb{R}^{dim}$, we calculate the similarity between the P and T based on matrix multiplication, which can be formulated as follows:

$$S_{P-T} = PT, \tag{3}$$

where $S_{P-T} \in \mathbb{R}^{n \times 1}$ is the similarity vector between the text and each patch.

Similar to patch-text contrast S_{P-T}, we calculate the similarity between the image feature at coarse-granularity $I \in \mathbb{R}^{dim}$ and text feature at fine-granularity $W \in \mathbb{R}^{m \times dim}$ by the matrix multiplication, which can be formulated as follows:

$$S_{I-W} = (WI)^{\top}, \tag{4}$$

where $S_{I-W} \in \mathbb{R}^{1 \times m}$ is the similarity vector between one image and each word in one sentence.

Attention-based Fusion on Similarity Matrix

To realize semantic information coverage from coarse-to-fine granularity and obtain instance-level similarity, we fuse the cross-modal contrast similarity of each granularity including Eq.(1), Eq.(2), Eq.(3), and Eq.(4), as the final similarity.

In order to distinguish the different importance of different patches and words on the fusion results, we propose an attention-based fusion module on a similarity matrix, dubbed AF, where scores in similarity will be given different weights during aggregation, realizing differential fusion.

Intra-granularity level fusion. Owing to the fine-granularity similarity matrix $S_{P-W} \in \mathbb{R}^{n \times m}$ that contains the similarity scores of n patches of one image and m words of one text, thus, we deploy attention operations on the S_{P-W} twice. The first attention is to obtain fine-granularity image-level and text-level similarity vectors, which can be formulated as:

$$S^{img}_{P-W} = \sum_{i=1}^{n} \frac{\exp\left(S_{P-W(i,\circledast)}/\tau\right)}{\sum_{j=1}^{n} \exp\left(S_{P-W(j,\circledast)}/\tau\right)} S_{P-W(i,\circledast)}, \tag{5}$$

$$S^{txt}_{P-W} = \sum_{i=1}^{m} \frac{\exp\left(S_{P-W(i,\circledast)}/\tau\right)}{\sum_{j=1}^{m} \exp\left(S_{P-W(j,\circledast)}/\tau\right)} S_{P-W(\circledast,i)}, \tag{6}$$

where the τ is the temperature hyper-parameter of softmax and the $\circledast$ denotes all content in the dimension. $S^{img}_{P-W} \in \mathbb{R}^{1 \times m}$ is the image-level similarity between the image and m words in the text. $S^{txt}_{P-W} \in \mathbb{R}^{n \times 1}$ is the text-level similarity between the text and n patches in the image. In order to further obtain the fine-granularity instance-level similarity score, we continue to carry out the second attention operation, which can be represented as:

$$S'^{img}_{P-W} = \sum_{i=1}^{m} \frac{\exp\left(S^{img}_{P-W(1,i)}/\tau\right)}{\sum_{j=1}^{m} \exp\left(S^{img}_{P-W(1,j)}/\tau\right)} S^{img}_{P-W(1,i)}, \tag{7}$$

$$S'^{txt}_{P-W} = \sum_{i=1}^{n} \frac{\exp\left(S^{txt}_{P-W(i,1)}/\tau\right)}{\sum_{j=1}^{n} \exp\left(S^{txt}_{P-W(j,1)}/\tau\right)} S^{txt}_{P-W(i,1)}, \tag{8}$$

where $S'^{img}_{P-W} \in \mathbb{R}^1$ and $S'^{txt}_{P-W} \in \mathbb{R}^1$ are the instance-level similarities. We take the average of similarities at instance-level as the final fine-granularity similarity:

$$S'_{P-W} = (S'^{img}_{P-W} + S'^{txt}_{P-W})/2. \tag{9}$$

Inter-granularity level fusion. We leverage the softmax function to obtain fusion weights. Then, we can aggregate these similarity scores according to the obtained weights, which can be described as:

$$S'_{P-T} = \sum_{i=1}^{n} \frac{\exp\left(S_{P-T(1,i)}/\tau\right)}{\sum_{j=1}^{n} \exp\left(S_{P-T(1,j)}/\tau\right)} S_{P-T(i,1)}, \tag{10}$$

$$S'_{I-W} = \sum_{i=1}^{m} \frac{\exp\left(S_{I-W(1,i)}/\tau\right)}{\sum_{j=1}^{m} \exp\left(S_{I-W(1,j)}/\tau\right)} S_{I-W(1,i)}, \tag{11}$$

Training and Inference

Similarity calculation. For a given pair $S(I_k, T_k)$ in $B = \{(I_k, T_k)\}_{k=1}^{N}$, the final similarity which contains multi-granularity contrastive scores can be described as follows:

$$S(I_k, T_k) = (S'_{P-W} + S_{I-T} + S'_{P-T} + S'_{I-W})/4. \tag{12}$$

Objective loss function. The InfoNCE loss function is utilized to pull the positive instances and push away the hard negative ones in a batch of B image-text pairs.

$$\mathcal{L}_{image2text} = -\frac{1}{B} \sum_{i=1}^{B} \log \frac{\exp\left(S(I_i, T_i)\right)}{\sum_{j=1}^{B} \exp\left(S(I_i, T_j)\right)}, \tag{13}$$

$$\mathcal{L}_{text2image} = -\frac{1}{B} \sum_{i=1}^{B} \log \frac{\exp\left(S(I_i, T_i)\right)}{\sum_{j=1}^{B} \exp\left(S(I_j, T_i)\right)}, \tag{14}$$

$$\mathcal{L} = \mathcal{L}_{image2text} + \mathcal{L}_{text2image}. \tag{15}$$

Experiments

In this section, we will manifest the results of experiments with corresponding analysis for TO-ReID. Concretely, 4 research questions (RQs) lead our discussions about experiments: **RQ1:** Is the overall performance of MGCC superior to the other SOTA baselines under occlusion? **RQ2:** Are the multi-granularity contrast modules effective and essential? **RQ3:** How does the token selection mechanism affect the MGCC model? **RQ4:** How to evaluate the quality of feature representation and retrieval ranking?

Experiment Settings

Datasets. We construct three occluded datasets via OGor, called Occluded-CUHK-PEDES, Occluded-ICFG-PEDES, and Occluded-RSTPReid, based on three existing T-ReID datasets.

Evaluation Metrics. For TO-ReID, we employ two evaluation metrics that are widely used in retrieval tasks, to measure the performance: Recall at K (R@K, higher is better), and mean Average Precision (mAP, higher is better). Meanwhile, we also adopt "Rsum" to measure the overall quality, which is defined as the sum of R@1, R@5, and R@10.

Comparison with SOTA Methods

To answer **RQ1**, we evaluate the performance of MGCC by comparing it with existing T-ReID models on three occluded datasets from two paradigms (*a.k.a.*, single-granularity and multi-granularity), the detailed results are shown in Table 2.

Performance Comparisons on Occluded-CUHK-PEDES. The MGCC can achieve comparable results to recent state-of-the-art methods, with 62.44%, 82.44%, and 88.52% on R@1, R@5, and R@10, respectively. Although our MGCC performs slightly worse than IRRA (Jiang and Ye 2023) on R@1 and R@5, it achieves optimal performance in Rsum index compared with other baselines, which reflects the overall robust retrieval quality of the proposed MGCC model.

Performance Comparisons on Occluded-ICFG-PEDES. Our MGCC model outperforms competitive candidates in terms of all metrics, achieving 59.28% R@1 accuracy, which significantly improved (+3.19% and +9.11%, respectively) in Rsum, compared with IRRA and CFine (Yan et al. 2022).

Performance Comparisons on Occluded-RSTPReid. The proposed MGCC dramatically surpasses the single-granularity feature learning paradigm Dual-Path (Zheng et al. 2020b) by 31.5% and 65.95% on R@1 and Rsum, respectively. Compared with the multi-granularity paradigm IRRA, MGCC also achieves great performance, with an increase of 1.2% and 3.67% on R@1 and Rsum, respectively.

Meanwhile, it is worth noting that the transformer-based powerful feature extraction backbones become more significant for better retrieval performance. In order to search for the most effective feature extraction backbones, we conduct ablation studies among ResNet50, LSTM, BERT, and CLIP, as shown in the "Baseline" row in Table 2. The comparison clearly demonstrates the effectiveness of multi-modal vision-language pre-training backbones.

Ablation Studies on Contrastive Modules

To answer **RQ2**, we evaluate proposed modules in Table 3.

We first investigate the influence of each independent contrastive modules. As a basic contrast, we could find that using only the coarse-granularity contrast S_{I-T} is powerful enough to outperform many SOTA baselines. Additionally, each independent contrast module can realize competitive results, indicating the effectiveness of our multi-granularity contrast consistency framework.

Based on the basic contrast S_{I-T}, the S_{P-W}, S_{I-W}, and S_{P-T} can enhance the model's performance by providing additional indirect matching information from different perspectives, thus realizing full coverage of semantics from coarse-to-fine. Finally, we equip all the contrastive modules, our MGCC can yield 62.44%, 59.28%, and 49.85% of R@1 on three datasets, respectively. Therefore, we conclude that all types of granularity contrastive consistency learning modules are effective and complementary to improve retrieval performance.

Influence of the Token Selection Mechanism

To answer **RQ3**, we analyze the influence from two aspects:

Robustness for feature representation. To further investigate the robustness on discriminate feature extraction, we visualize the token selection process on images and texts in *Appendix (§E)*. It clearly proves that the token selection mechanism can enhance the effectiveness of the MGCC model by eliminating uninformative tokens (occlusions and backgrounds in images, *etc.*), thus promoting the model to focus on the most discriminative part.

Improvement of training efficiency. As shown in Table 4, the no-selection baseline (*a.k.a.*, (ρ_i, ρ_t) = (1.0, 1.0)) and the best trade-off experiments ((ρ_i, ρ_t) = (0.3, 0.4) on Occluded-CUHK-PEDES, (ρ_i, ρ_t) = (0.4, 0.5) on Occluded-ICFG-PEDES, and (ρ_i, ρ_t) = (0.5, 0.5) on Occluded-RSTPReid) are typically compared to make a conclusion: After equipping with the token selection mechanism, the computational memories (M) can be reduced by 16.04%∼16.83% and the inference time (T) can be accelerated by 28.10% ∼49.58%, with a slight influence of R@1 performance (-0.49%∼+3.10%). *Detailed ablation on ρ_i and ρ_t are shown in Appendix (§F).*

Qualitative Results

To answer **RQ4**, we illustrate from following two aspects:

Feature representation visualization. We adopt the t-SNE (Van der Maaten and Hinton 2008) to visualize the difference of feature representations before and after alignment, which aims at showing the model's effectiveness in narrowing the modality gap. *Detailed t-SNE visualization processes are shown in Appendix (§G).*

Retrieval ranking visualization. As shown in Figure 4, for each text query, the top-10 matches are displayed. On one hand, the orange highlighted words represent the selected high-informative fine-grained tokens, which dominants the retrieval process. On the other hand, although the candidate images are partially occluded, our MGCC model can still successfully retrieve the correct pedestrian, thereby showcasing the outstanding retrieval capability of MGCC.

Method (Type)	Image-Text Encoder	Occluded-CUHK-PEDES					Occluded-ICFG-PEDES					Occluded-RSTPReid				
		R@1	R@5	R@10	mAP	Rsum	R@1	R@5	R@10	mAP	Rsum	R@1	R@5	R@10	mAP	Rsum
DPath (S) (2020b)	RN50-RN50	33.69	60.48	71.30	30.12	165.47	29.03	53.29	64.50	14.03	146.82	18.35	42.65	56.35	14.80	117.35
CMPM (S) (2018)	RN50-LSTM	40.57	65.38	70.54	32.60	176.49	35.43	57.22	69.43	-	162.08	-	-	-	-	-
TBPS (S) (2021)	RN50-BERT	58.85	79.34	86.05	52.64	224.24	-	-	-	-	-	-	-	-	-	-
TIPCB (S) (2022)	RN50-BERT	59.91	80.22	86.66	-	226.79	-	-	-	-	-	-	-	-	-	-
ViTAA (S) (2020b)	RN50-LSTM	55.97	75.84	83.52	-	215.33	-	-	-	-	-	-	-	-	-	-
NAFS (M) (2021)	RN50-BERT	56.85	75.39	**90.27**	49.58	222.51	-	-	-	-	-	-	-	-	-	-
SAF (M) (2022)	ViT-BERT	59.47	79.65	84.90	50.27	224.02	-	-	-	-	-	-	-	-	-	-
SSAN (M) (2021)	RN50-LSTM	59.76	78.98	85.48	**59.35**	224.22	53.01	71.83	78.90	30.00	203.74	39.85	66.45	76.50	29.85	182.80
CFine (M) (2022)	CLIP-BERT	56.24	76.14	83.80	47.72	216.18	55.55	75.14	82.55	30.94	213.24	40.50	63.40	73.40	31.75	177.30
LGUR (M) (2022)	DeiT-BERT	59.82	79.30	86.80	55.81	225.92	52.91	70.41	77.01	30.46	200.33	47.40	71.80	80.60	35.07	199.80
IRRA (M) (2023)	CLIP-CLIP	**64.41**	**82.63**	85.43	57.56	232.47	57.24	78.07	83.85	30.42	219.16	48.65	**75.43**	80.50	**39.48**	204.58
Baseline1 (M)	RN50-LSTM	52.80	67.91	75.80	49.21	196.51	52.40	73.30	80.80	30.22	206.50	31.50	46.55	55.95	22.23	134.00
Baseline2 (M)	CLIP-BERT	61.00	80.51	86.76	54.74	228.27	55.68	75.50	82.35	32.75	213.53	41.58	66.75	75.65	31.86	183.98
MGCC (M)	CLIP-CLIP	62.44	82.44	88.52	54.18	**233.40**	**59.28**	**78.32**	**84.75**	**33.30**	**222.35**	**49.85**	74.95	**83.45**	38.48	**208.25**

Table 2: Performance comparisons on three occluded datasets. "S" and "M" in "Type" stand for Single/Multi-granularity paradigm.

Variants of MGCC				Datasets														
intra-granularity parallel contrast		inter-granularity cross contrast		Occluded-CUHK-PEDES					Occluded-ICFG-PEDES					Occluded-RSTPReid				
S_{I-T}	S_{P-W}	S_{P-T}	S_{I-W}	R@1	R@5	R@10	mAP	Rsum	R@1	R@5	R@10	mAP	Rsum	R@1	R@5	R@10	mAP	Rsum
✓				57.33	77.70	84.45	49.13	219.48	57.88	76.36	82.72	31.57	216.96	41.30	64.60	74.95	31.52	180.85
	✓			59.76	80.36	87.52	52.46	227.64	55.26	76.18	83.09	31.76	214.53	43.00	69.40	79.15	33.89	191.55
		✓		59.55	80.75	87.41	52.76	227.71	56.57	76.83	84.05	32.13	217.45	41.05	69.40	78.45	33.23	188.90
			✓	49.69	70.42	77.71	41.29	197.82	49.73	70.49	77.69	24.88	197.91	33.05	55.25	64.45	24.83	152.75
✓	✓			62.23	**82.68**	**88.86**	**54.55**	**233.77**	58.52	77.99	**84.76**	32.77	221.27	48.80	72.90	82.90	37.29	204.60
		✓	✓	61.92	82.47	88.65	54.03	233.04	58.60	77.85	84.51	32.95	220.96	46.55	72.95	82.20	36.12	201.70
✓	✓	✓	✓	**62.44**	82.44	88.52	54.18	233.40	**59.28**	**78.32**	84.75	**33.30**	**222.35**	**49.85**	**74.95**	**83.45**	**38.48**	**208.25**

Table 3: A series of ablation studies on three occluded datasets to investigate effects of different contrastive modules.

Dataset	(ρ_i, ρ_t)	R@1↑	mAP↑	Rsum↑	T↓	M ↓
Occluded -CUHK-PEDES	(1.0, 1.0)	**62.75**	**54.53**	**234.49**	3.63	15.27
	(0.3, 0.4)	62.44	54.18	233.40	**2.61**	**12.70**
Occluded -ICFG-PEDES	(1.0, 1.0)	58.80	33.12	221.45	8.33	15.27
	(0.2, 0.4)	**59.28**	**33.30**	**222.35**	**4.40**	**12.70**
Occluded -RSTPReid	(1.0, 1.0)	48.35	37.00	202.40	3.94	15.27
	(0.5, 0.5)	**49.85**	**38.48**	**208.25**	**3.24**	**12.82**

Table 4: Comparison with different token selection ratios. "↑" means the higher, the better; "↓" means the lower, the better.

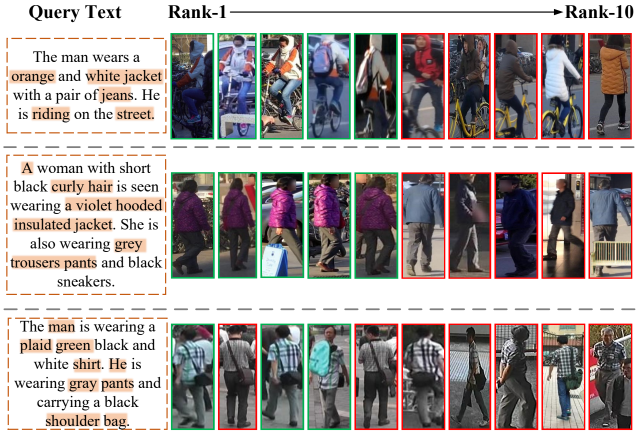

Query Text
Rank-1
Rank-10
The man wears a orange and white jacket with a pair of jeans. He is riding on the street.
A woman with short black curly hair is seen wearing a violet hooded insulated jacket. She is also wearing grey trousers pants and black sneakers.
The man is wearing a plaid green black and white shirt. He is wearing gray pants and carrying a black shoulder bag.

Figure 4: Retrieval ranking visualization. Matched and mismatched images are marked with green and red, respectively.

Conclusion

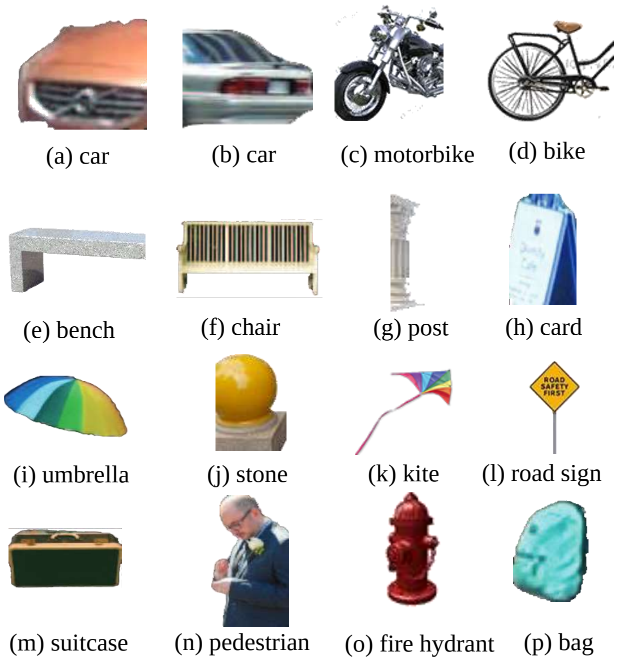
In this paper, we make the first attempt to tackle a complex and challenging problem, TO-ReID. To handle this tricky issue, we design an OGor to generate occluded persons for simulating the real-world scenario. Meanwhile, a novel MGCC framework is proposed, to narrow the semantic gap and modality gap. Experimental results show the effectiveness and superiority of our proposal.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 62072465, No. 62302144) and the Science and Technology Innovation Program of Hunan Province (No. 2022RC3061, No. 2023RC3027).

§A Occlusion Instances in OIL

As shown in Figure 5, we selected 15 occlusion instance samples for display (one sample for each category), which are the main components of our Occusion Instance Library (OIL).



(a) car (b) car (c) motorbike (d) bike

(e) bench (f) chair (g) post (h) card

(i) umbrella (j) stone (k) kite (l) road sign

(m) suitcase (n) pedestrian (o) fire hydrant (p) bag

Figure 5: occluded instance samples from OIL.

§B Occlusion generation algorithm

The detailed occlusion generation algorithm is described in Algorithm 1.

For each selected image $X_i \in \mathbb{R}^{H\times W\times C}$, where H, W, and C represent the height, width, and channel dimensions of the holistic image, we paste the occlusion masks onto the corresponding regions of the image.

Specifically, we follow these steps:

(1) Randomly select an occlusion mask instance sample $x_o \in \mathbb{R}^{h\times w\times C}$ from OIL.

(2) Define the new occlusion area S_o as 10%$\sim$60% of the holistic image area $S = H \times W$, while keeping the original aspect ratio $\gamma = \frac{h}{w}$ unchanged.

(3) Perform a resize operation $\varphi(x_o, \gamma)$ to obtain the resized occlusion instance x_o^{new}, which has a height h_o of $\sqrt{S_0 \times \gamma}$ and a width w_o of $\sqrt{S_0/\gamma}$.

(4) Determine the location for augmentation. If $x_o \in O_U$, we set the top-left corner coordinates to $(0, 0)$. If $x_o \in O_M$, we set the top-left corner coordinates to $(range(0, \frac{H}{2} - h_o), range(0, W))$. If $x_o \in O_B$, we set the top-left corner coordinates of the occlusion instance to $(H - h_o, 0)$.

(5) Paste the resized occlusion instance onto the holistic image to generate an occluded image. Following this process, we will obtain occluded copies of every selected holistic image.

Algorithm 1: The Occlusion Generation Algorithm

```
Input: Holistic image $X_i \in \mathbb{R}^{\mathbf{H\times W\times C}}$ with the area
    $S = H * W$; Generated OIL $\{O_U, O_M, O_B\}$.
Output: Artificial occluded image $X_i^o \in \mathbb{R}^{\mathbf{H\times W\times C}}$.
while True do
    1. Occlusion instance $x_o \in \mathbb{R}^{\mathbf{h\times w\times C}}$;
      // Randomly choose an occlusion instance
      from OIL;
    2. $S_o = \delta \times S$, $\delta \sim \mathbf{U}(0.1, 0.6)$;
      // Randomly generate the occluded area $S_o$, $\delta$
      is a scaling factor, $\mathbf{U}$ denotes the uniform
      distribution;
    3. $\gamma = \frac{h}{w}$ ;
      // keep the aspect ratio of $x_o$ unchanged;
    4. $x_o^{new} = \varphi(x_o, \gamma)$,    $x_o^{new} \in \mathbb{R}^{\mathbf{h_o\times w_o\times C}}$,
      where $h_o = \sqrt{S_o * \gamma}$, $w_o = \sqrt{S_o/\gamma}$;
      // Perform a resize operation for the occlusion
      instance;
    5.Define the upper left corner coordinate of
      $x_o^{new}$ : $(h_l, w_l)$;
      // Determine the location for the occlusion
      instance;
    if $x_o \in O_U$ then
        $(h_l, w_l) = (0, 0)$;
    if $x_o \in O_M$ then
        $(h_l, w_l) =$
          $(range(0, \frac{H}{2} - h_o), range(0, W))$;
    if $x_o \in O_B$ then
        $(h_l, w_l) = (H - h_o, 0)$;
    if $(h_l + h_o) < H$ and $(w_l + w_o) < W$ then
        $x_o^{new} = (h_l, w_l, h_l + h_o, w_l + w_o)$ ;
        // Diagonal coordinates of the occlusion area;
        $X_i^o = X_i + x_o^{new}$ ;
        // Paste the occlusion area onto the holistic
          image to generate an occluded image;
        return $X_i^o$.
```

§C Datasets

The **Occluded-CUHK-PEDES** dataset contains 13003 identities with 40206 images and 80412 textual descriptions. Firstly, we use the OGor to preprocess the images, then follow the official splitting guideline: 11003 identities with 34054 images and 68108 descriptions are used for training, while 1000 identities with 3078 images and 1000 identities with 3074 images are for validation and testing, respectively.

The **Occluded-ICFG-PEDES** dataset consists of 54522 pedestrian images and 4102 identities with 1 textual description per image. Firstly, we employ the OGor to preprocess the images, then follow the common splitting guideline for deploying the experiments: 3102 identities with 34674 image-text pairs are used for training and 19848 image-text pairs of 1000 identities are for testing.

The **Occluded-RSTPReid** dataset includes 41010 textual descriptions and 20505 images of 4101 pedestrian persons. That is, per identity has 5 corresponding images caught by distinct cameras, and each image is manually annotated with 2 textual descriptions.

§D Implementation Details

The experiments are conducted with one NVIDIA Tesla A100 40GB GPU using the PyTorch platform. Similar to (Luo et al. 2022; Ma et al. 2022), both the image encoder and text encoder are adopted from the CLIP checkpoints with ViT-B/32. The framework is optimized by Adam optimizer (Kingma and Ba 2014) that the initial learning rate for text encoder and image encoder is 10^{-5}, the initial learning rate for other modules is 10^{-4}, and decay the learning rate using a cosine schedule strategy (Loshchilov and Hutter 2016). Meanwhile, we set batch size, and training epoch to 128 and 60. The patch number n is defined as 49 and the word token number m is limited to 25. For feature representations of all granularity, the embedded dimension *dim* of image-text modalities is 768. For the token selection ratio, we choose the best $(\rho_i, \rho_t) = (0.3, 0.4)$ on Occluded-CUHK-PEDES, $(\rho_i, \rho_t) = (0.4, 0.5)$ on Occluded-ICFG-PEDES, and $(\rho_i, \rho_t) = (0.5, 0.5)$ on Occluded-RSTPReid. Detailed ablation studies of (ρ_i, ρ_t) are shown in *§F*. For the AF module, we set the hyper-parameter τ to 0.01 on all datasets by default. Detailed ablation studies of the hyper-parameter τ and concrete AF strategy are shown in *§H*.

§E Visualization of the token selection process

The visualization results of the fine-grained token selection mechanism are shown in Figure 6. To demonstrate the effectiveness of the token selection mechanism in various occlusion scenarios, the following subfigures present the results. These results verify that this mechanism can effectively mitigate the impact of occlusions or background noises, allowing for the extraction of robust visual features.

From the following four examples, we can draw some significant observations:

- It clearly proves that our token selection mechanism can not only drop the uninformative tokens including occlusions and backgrounds but also prompt the model to focus on the most discriminative part.
- Additionally, we present the visualization results of word token selection. As depicted in the figure, our model can accurately focus on the word tokens containing rich semantic information with the help of the token selection mechanism.
- In short, we leverage the attention map between the global [CLS] token and every local token to select semantically rich fine-grained features, which not only reduces redundant features and memory overhead but also improves inference efficiency.

§F Ablation Studies of the token selection ratio ρ_i and ρ_t

Since the differences between datasets, we conduct a series of ablation studies about the token selection ratios. There are two cases for ρ_i and ρ_t: if the ratio is too small, discriminative features at the fine-granularity level will be ignored; if the ratio is too large, it will produce feature redundancy and background noise, resulting in decayed performance and extra memory overhead. Specifically, we manually adjust the ratios between the image token selection ratio ρ_i and the text token selection ratio ρ_t with a step size of 0.1 (from 0.1 to 0.5), to evaluate the impact of the trade-off token selection ratio. Detailed experiment results are shown in Table 5, for each dataset, we leverage the **Control Variable Method** to implement ablation experiments and assess the models from the perspective of R@K, mAP, inference efficiency, and memory overhead. From Table 5, we can draw the following observations:

- The performance comparison under different token selection ratios is shown in Table 5, which reflects that our fine-granularity token selection mechanism consistently performs well within a specific range on the three occluded datasets.
- For the token selection ratio, we choose the $(\rho_i, \rho_t) = (0.3, 0.4)$ on Occluded-CUHK-PEDES, $(\rho_i, \rho_t) = (0.4, 0.5)$ on Occluded-ICFG-PEDES, and $(\rho_i, \rho_t) = (0.5, 0.5)$ on Occluded-RSTPReid, respectively. The parameters mentioned above (bold parts in Table 5) can not only promote a pick R@K but also greatly save the memory overhead and accelerate the inference process at the same time.
- In general, equipping the token selection mechanism, the computational memories can be reduced by $16.04\% \sim 16.83\%$, and the inference time can be accelerated by $28.10\% \sim 49.58\%$, with a slight influence of R@1 performance $(-0.49\% \sim +3.10\%)$.

§G Feature Representation Visualization of T-SNE

We adopt the t-SNE to embed the feature representations into the semantic common space of visual-textual for visually showing the effectiveness. As shown in Figure 8 (a), one can see that in the common space, the inter- and intra-identity samples from both image modality and text modality are hard to be distinguished. From Figure 8 (b), we can

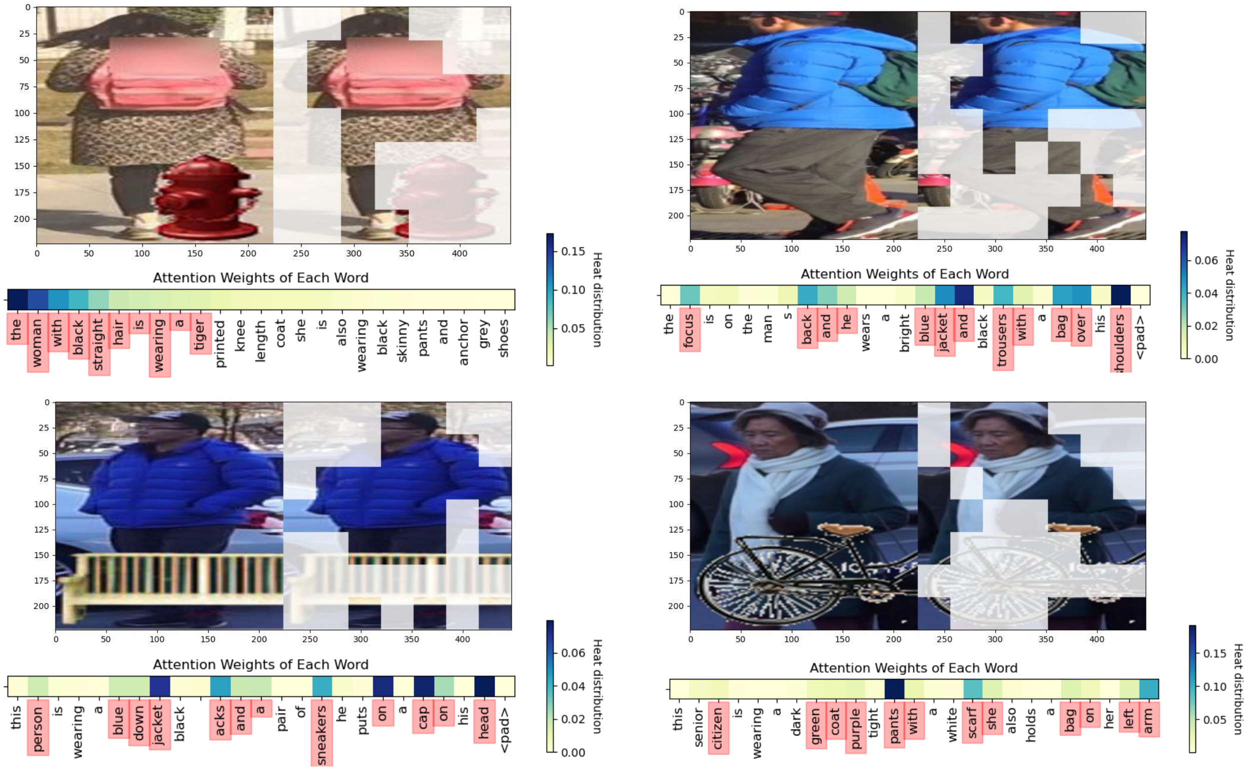


Figure 6: Visualization of the fine-granularity token selection process. Each subfigure has three components, which respectively represent the raw occluded image in the top left, effective occluded images after token selection in the top right, and the attention-weight visualization of each word in the text description.

Datasets	Evaluation Metrics	$\rho_t = 1.0$	$\rho_t = 0.1$					$\rho_t = 0.2$					$\rho_t = 0.3$					$\rho_t = 0.4$					$\rho_t = 0.5$				
		$\rho_i = 1.0$	$\rho_i = 0.1$	$\rho_i = 0.2$	$\rho_i = 0.3$	$\rho_i = 0.4$	$\rho_i = 0.5$	$\rho_i = 0.1$	$\rho_i = 0.2$	$\rho_i = 0.3$	$\rho_i = 0.4$	$\rho_i = 0.5$	$\rho_i = 0.1$	$\rho_i = 0.2$	$\rho_i = 0.3$	$\rho_i = 0.4$	$\rho_i = 0.5$	$\rho_i = 0.1$	$\rho_i = 0.2$	$\rho_i = 0.3$	$\rho_i = 0.4$	$\rho_i = 0.5$	$\rho_i = 0.1$	$\rho_i = 0.2$	$\rho_i = 0.3$	$\rho_i = 0.4$	$\rho_i = 0.5$
Occluded -CUHK-PEDES	R@1 ↑	62.75	61.91	61.50	61.03	60.75	60.62	61.70	61.21	61.01	61.31	61.50	60.92	61.45	62.02	61.57	60.48	61.05	61.18	**62.44**	61.86	61.96	61.48	61.92	61.68	62.04	61.21
	R@5 ↑	82.83	81.84	82.33	82.68	82.10	82.47	81.76	82.39	81.87	81.40	82.34	82.18	81.50	82.23	82.42	81.81	81.73	82.29	**82.44**	82.23	82.21	82.23	81.95	81.74	81.81	81.69
	R@10 ↑	88.91	88.55	88.61	88.53	88.21	88.87	87.61	88.84	88.42	87.90	88.95	88.56	88.74	88.27	88.61	88.16	88.40	88.74	**88.52**	88.89	88.68	88.55	88.52	88.34	88.42	88.68
	mAP ↑	54.53	53.42	53.48	53.46	53.23	53.41	53.65	53.73	53.40	53.55	53.61	53.42	53.46	53.77	53.91	53.03	52.94	53.71	**54.18**	53.74	54.17	53.65	54.04	53.81	53.95	53.26
	Rsum ↑	234.49	232.30	232.44	232.24	231.06	231.96	231.07	232.44	231.30	230.61	232.79	231.66	231.69	232.52	232.60	230.45	231.18	232.21	**233.40**	232.98	232.85	232.26	232.39	231.76	232.27	231.58
	T (ms) ↓	3.63	2.69	2.76	2.60	2.65	3.08	2.53	2.73	3.01	2.75	3.39	2.69	2.55	2.79	3.20	2.93	2.91	2.61	**2.61**	2.85	2.69	2.79	2.79	2.84	2.70	2.83
	M (GB) ↓	15.27	12.68	12.69	12.70	12.70	12.71	12.68	12.69	12.70	12.71	12.71	12.69	12.69	12.70	12.72	12.75	12.69	12.70	**12.70**	12.75	12.79	12.69	12.70	12.73	12.78	12.82
Occluded -ICFG-PEDES	R@1 ↑	58.80	57.12	57.07	57.18	57.65	57.33	57.94	57.96	57.89	58.41	57.34	57.95	57.88	58.68	58.22	57.62	57.83	58.76	58.52	58.08	58.66	58.37	57.44	58.15	**59.28**	58.43
	R@5 ↑	78.03	77.25	77.11	77.43	77.41	77.10	77.57	77.65	77.36	78.16	77.78	78.15	77.72	77.98	78.61	78.21	77.55	77.75	77.91	77.77	78.31	78.64	77.30	78.12	**78.32**	78.21
	R@10 ↑	84.62	84.34	84.04	84.05	84.13	83.86	84.32	84.41	84.59	84.78	84.45	84.63	84.24	84.68	84.85	85.14	84.15	84.67	84.69	84.39	84.91	84.97	84.25	84.60	**84.75**	84.75
	mAP ↑	33.12	32.34	32.43	32.38	32.66	32.22	32.71	32.65	32.59	33.02	32.63	32.92	32.50	33.12	33.14	33.24	32.71	32.99	32.85	32.86	33.30	32.99	32.61	32.92	**33.30**	33.16
	Rsum ↑	221.45	218.71	218.22	218.66	219.19	218.29	219.83	220.02	219.84	221.35	219.57	220.73	219.84	221.34	221.68	220.97	219.53	221.18	221.12	220.24	221.88	221.98	218.99	220.87	**222.35**	221.39
	T (ms) ↓	8.33	4.63	4.67	5.13	4.20	4.24	4.32	4.25	4.76	4.63	4.72	5.05	4.82	4.58	4.67	4.85	4.59	4.58	4.26	4.44	4.52	4.67	5.86	4.47	**4.40**	4.32
	M (GB) ↓	15.27	12.68	12.69	12.70	12.70	12.71	12.68	12.69	12.70	12.71	12.71	12.69	12.69	12.70	12.72	12.75	12.69	12.78	12.70	12.75	12.79	12.69	12.70	12.73	**12.70**	12.82
Occluded -RSTPReid	R@1 ↑	48.35	44.95	45.65	43.60	44.90	46.10	43.25	43.45	46.45	46.05	44.95	44.85	44.50	48.30	44.95	46.20	45.00	45.50	46.50	48.05	46.45	47.10	45.50	45.80	46.25	**49.85**
	R@5 ↑	71.90	71.25	71.55	70.00	71.05	71.35	71.20	70.55	71.90	71.05	71.30	71.95	71.25	72.30	71.15	72.40	71.60	70.50	73.00	72.00	72.10	72.30	71.85	71.60	72.05	**74.95**
	R@10 ↑	82.15	81.65	80.70	80.50	81.60	82.30	81.95	80.55	82.20	81.25	81.80	81.90	81.05	81.20	81.00	81.10	80.60	79.35	82.15	81.90	81.00	81.95	82.20	80.90	80.85	**83.45**
	mAP ↑	37.00	35.61	35.35	35.07	36.18	35.95	34.78	34.33	36.36	35.89	35.64	35.90	35.47	36.51	35.87	35.70	35.53	35.55	36.94	37.00	36.23	36.66	35.87	36.37	36.08	**38.48**
	Rsum ↑	202.40	197.85	197.90	194.10	197.55	199.75	196.40	194.55	200.55	198.35	198.05	198.70	196.80	201.80	197.10	199.70	197.20	195.35	201.65	201.95	199.55	201.35	199.55	198.30	199.15	**208.25**
	T (ms) ↓	3.94	3.39	3.31	3.60	3.42	3.44	3.29	3.04	3.49	3.07	3.39	3.61	3.69	3.31	3.63	2.86	3.16	3.48	2.88	2.95	3.60	2.68	2.92	2.92	3.09	**3.24**
	M (GB) ↓	15.27	12.68	12.69	12.70	12.70	12.71	12.68	12.69	12.70	12.71	12.71	12.69	12.69	12.70	12.72	12.75	12.69	12.70	12.70	12.75	12.79	12.69	12.70	12.73	12.78	**12.82**

Table 5: Ablation Studies of the token selection ratios. “↑” means the higher, the better; “↓” means the lower, the better; “T” stands for Inference Time, and “M” stands for Computational Memories.

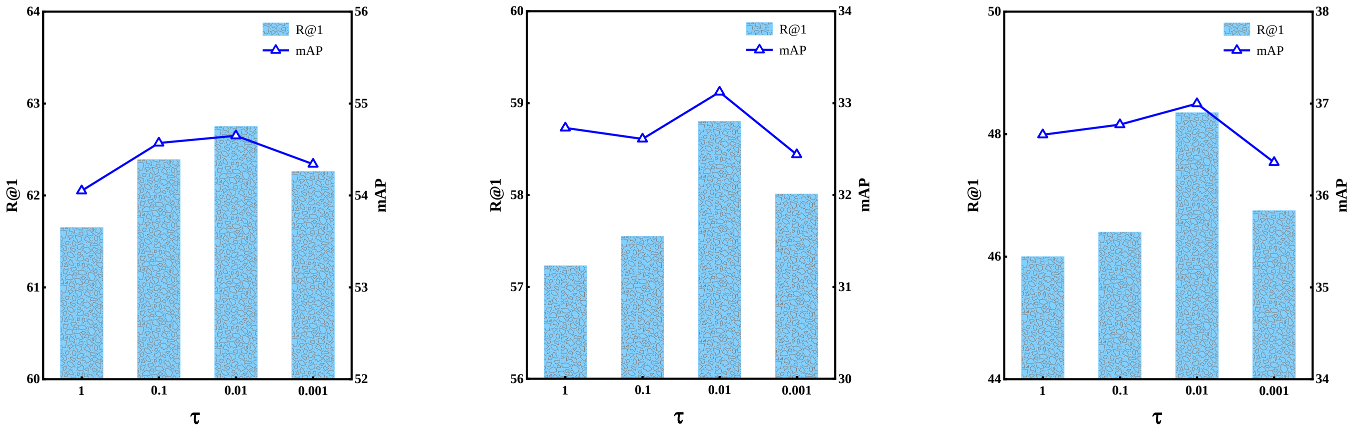
see that our MGCC can learn the feature representations of image-text pairs with strong discrimination, namely, by dividing the feature representations of different modalities into several semantically distinct clusters visually. As a result, this means that our proposal can effectively bridge the modality gap between image and text.

§H: Fusion Strategy AF and Hyper-parameter τ

At the same time, we assess the effect of distinct fusion strategies and the temperature hyper-parameter τ on our proposed MGCC.

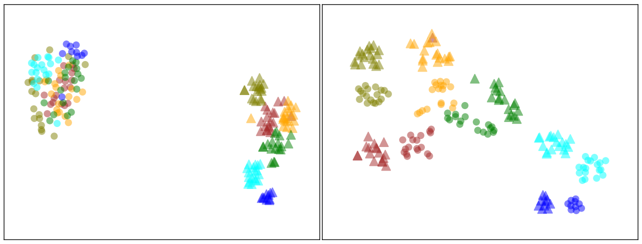
Fusion Strategies	Occluded-CUHK-PEDES				Occluded-ICFG-PEDES				Occluded-RSTPReid			
	R@1	R@5	R@10	mAP	R@1	R@5	R@10	mAP	R@1	R@5	R@10	mAP
Mean-Mean	60.25	80.75	87.80	52.89	56.88	77.10	84.21	31.81	45.40	70.90	81.50	34.87
Mean-Max	60.59	82.12	**88.65**	53.13	56.73	77.44	84.29	32.14	47.65	71.65	80.40	36.60
Max-Mean	60.83	81.94	88.11	53.30	57.59	77.30	84.52	32.73	45.75	71.30	80.65	35.40
Max-Max	61.44	82.47	88.53	53.61	58.21	77.82	84.40	33.14	45.30	71.25	81.25	36.11
Our AF	**62.44**	**82.44**	88.52	**54.18**	**59.28**	**78.32**	**84.75**	**33.30**	**49.85**	**74.95**	**83.45**	**38.48**

Table 6: Retrieval performance with different fusion strategies for similarity matrices on Occluded-CUHK-PEDES, Occluded-ICFG-PEDES, and Occluded-RSTPReid.



(a) Occluded-CUHK-PEDES (b) Occluded-ICFG-PEDES (c) Occluded-RSTPReid

Figure 7: The sensitivity of the temperature factor τ.



(a) Features before alignment (b) Features after alignment

Figure 8: Feature visualization of common representation space under pre-training and the MGCC, for the six category data from Occluded-ICFG-PEDES by using the t-SNE. The same color indicates relevant semantics, the shapes represent different modalities.

As shown in Table 6, we conduct experiments with different fusion strategies including Mean-Mean, Mean-Max, Max-Mean, Max-Max, and our proposed AF module on three occluded datasets. For the conventional fusion strategies, the Mean-Mean strategy has the worst performance. This may be because the Mean-Mean strategy treats all similarity scores equally during aggregating, hard realizing differential fusion to ease noise and redundant feature representations. The other variants of Mean-Mean including Max-Max, Max-Mean, and Mean-Max, they can yield moderate results since these strategies employ the highest similarity during aggregation, thus eliminating noise and redundancy to a certain extent. Unfortunately, these strategies only leverage top-1 similarity score that ignores the influence of hard negative samples, failing to realize the best mAP. To handle this problem, we propose the AF module to enable differential fusion of all similarity scores based on attention assignment, which can tackle the noise and redundancy. It can be seen from the results, our AF can yield leading performance in all indexes on three datasets.

Since the temperature parameter τ controls the penalty intensity of attention, which is sensitive in the softmax function. As a result, we also discuss the influence of hyper-parameter τ. From Figure 7, we tried to manually adjust the τ at a step size of 10^{-1}, such as 0.1 and 0.01, and 0.001, to observe the effects. The results show that the performance (for R@1 and mAP indicators) of our model first improves before reaching the saturation point (*i.e.*, τ = 0.01) with the increase of τ, and then begins to decline. The main reason is that the similarity scores are very sensitive to τ, and whether it is too large or too small will have a harmful effect on the performance. Therefore, the hyper-parameter τ = 0.01 is adopted by default in our work.

References

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Cao, M.; Yang, T.; Weng, J.; Zhang, C.; Wang, J.; and Zou, Y. 2022. Locvtp: Video-text pre-training for temporal localization. In *Proc. of ECCV*, 38–56. Springer.

Chen, P.; Liu, W.; Dai, P.; Liu, J.; Ye, Q.; Xu, M.; Chen, Q.; and Ji, R. 2021a. Occlude them all: Occlusion-aware attention network for occluded person re-id. In *Proc. of ICCV*, 11833–11842.

Chen, T.; Xu, C.; and Luo, J. 2018. Improving text-based person search by spatial matching and adaptive threshold. In *Proc. of WACV*, 1879–1887.

Chen, Y.; Huang, R.; Chang, H.; Tan, C.; Xue, T.; and Ma, B. 2021b. Cross-modal knowledge adaptation for language-based person search. *IEEE Transactions on Image Processing*, 30: 4057–4069.

Chen, Y.; Zhang, G.; Lu, Y.; Wang, Z.; and Zheng, Y. 2022. TIPCB: A simple but effective part-based convolutional baseline for text-based person search. *Neurocomputing*, 494: 171–181.

Ding, C.; Wang, K.; Wang, P.; and Tao, D. 2020. Multi-task learning with coarse priors for robust part-aware person re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3): 1474–1488.

Ding, Z.; Ding, C.; Shao, Z.; and Tao, D. 2021. Semantically self-aligned network for text-to-image part-aware person re-identification. *ArXiv:2107.12666*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *Proc. of ICLR*.

Fang, H.; Xiong, P.; Xu, L.; and Chen, Y. 2021. Clip2video: Mastering video-text retrieval via image clip. *arXiv preprint arXiv:2106.11097*.

Farooq, A.; Awais, M.; Kittler, J.; and Khalid, S. S. 2021. AXM-Net: Implicit Cross-Modal Feature Alignment for Person Re-identification. In *Proc. of AAAI*.

Gao, C.; Cai, G.; Jiang, X.; Zheng, F.; Zhang, J.; Gong, Y.; Peng, P.; Guo, X.; and Sun, X. 2021. Contextual non-local alignment over full-scale representation for text-based person search. *arXiv preprint arXiv:2101.03036*.

Gao, S.; Wang, J.; Lu, H.; and Liu, Z. 2020. Pose-guided visible part matching for occluded person reid. In *Proc. of CVPR*, 11744–11752.

Ge, J.; Gao, G.; and Liu, Z. 2019. Visual-textual association with hardest and semi-hard negative pairs mining for person search. *arXiv preprint arXiv:1912.03083*.

Han, X.; He, S.; Zhang, L.; and Xiang, T. 2021. Text-Based Person Search with Limited Data. In *Proc. of BMVC*.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask r-cnn. In *Proc. of ICCV*, 2961–2969.

He, L.; Liang, J.; Li, H.; and Sun, Z. 2018. Deep spatial feature reconstruction for partial person re-identification: Alignment-free approach. In *Proc. of CVPR*, 7073–7082.

He, L.; and Liu, W. 2020. Guided saliency feature learning for person re-identification in crowded scenes. In *Proc. of ECCV*, 357–373. Springer.

He, L.; Wang, Y.; Liu, W.; Zhao, H.; Sun, Z.; and Feng, J. 2019. Foreground-aware pyramid reconstruction for alignment-free occluded person re-identification. In *Proc. of ICCV*, 8450–8459.

Jiang, D.; and Ye, M. 2023. Cross-Modal Implicit Relation Reasoning and Aligning for Text-to-Image Person Retrieval. In *Proc. of CVPR*, 2787–2797.

Jing, Y.; Si, C.; Wang, J.; Wang, W.; Wang, L.; and Tan, T. 2020. Pose-guided multi-granularity attention network for text-based person search. In *Proc. of AAAI*, volume 34, 11189–11196.

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

Li, S.; Cao, M.; and Zhang, M. 2022. Learning semantic-aligned feature representation for text-based person search. In *Proc. of ICASSP*, 2724–2728. IEEE.

Li, S.; Xiao, T.; Li, H.; Zhou, B.; Yue, D.; and Wang, X. 2017. Person search with natural language description. In *Proc. of CVPR*, 1970–1979.

Liu, J.; Zha, Z.-J.; Hong, R.; Wang, M.; and Zhang, Y. 2019. Deep adversarial graph attention convolution network for text-based person search. In *Proc. of ACM MM*, 665–673.

Loshchilov, I.; and Hutter, F. 2016. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*.

Luo, H.; Ji, L.; Zhong, M.; Chen, Y.; Lei, W.; Duan, N.; and Li, T. 2022. CLIP4Clip: An empirical study of CLIP for end to end video clip retrieval and captioning. *Neurocomputing*, 508: 293–304.

Ma, W.; Wu, X.; Zhao, S.; Zhou, T.; Guo, D.; Gu, L.; Cai, Z.; and Wang, M. 2023. FedSH: Towards Privacy-preserving Text-based Person Re-Identification. *IEEE Transactions on Multimedia*, 1–13.

Ma, Y.; Xu, G.; Sun, X.; Yan, M.; Zhang, J.; and Ji, R. 2022. X-CLIP: End-to-End Multi-grained Contrastive Learning for Video-Text Retrieval. In *Proc. of ACM MM*, 638–647.

Miao, J.; Wu, Y.; Liu, P.; Ding, Y.; and Yang, Y. 2019. Pose-guided feature alignment for occluded person re-identification. In *Proc. of ICCV*, 542–551.

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

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *Proc. of ICML*, 8748–8763. PMLR.

Shao, Z.; Zhang, X.; Fang, M.; Lin, Z.; Wang, J.; and Ding, C. 2022. Learning granularity-unified representations for text-to-image person re-identification. In *Proc. of ACM MM*, 5566–5574.

Shu, X.; Wen, W.; Wu, H.; Chen, K.; Song, Y.; Qiao, R.; Ren, B.; and Wang, X. 2022. See finer, see more: Implicit modality alignment for text-based person retrieval. In *Proc. of ECCV*, 624–641. Springer.

Sun, Y.; Zheng, L.; Yang, Y.; Tian, Q.; and Wang, S. 2018. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *Proc. of ECCV*, 480–496.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(11).

Wang, C.; Luo, Z.; Lin, Y.; and Li, S. 2021a. Text-based Person Search via Multi-Granularity Embedding Learning. In *Proc. of IJCAI*, 1068–1074.

Wang, G.; Yang, S.; Liu, H.; Wang, Z.; Yang, Y.; Wang, S.; Yu, G.; Zhou, E.; and Sun, J. 2020a. High-order information matters: Learning relation and topology for occluded person re-identification. In *Proc. of CVPR*, 6449–6458.

Wang, K.; Wang, P.; Ding, C.; and Tao, D. 2021b. Batch coherence-driven network for part-aware person re-identification. *IEEE Transactions on Image Processing*, 30: 3405–3418.

Wang, Y.; Bo, C.; Wang, D.; Wang, S.; Qi, Y.; and Lu, H. 2019. Language person search with mutually connected classification loss. In *Proc. of ICASSP*, 2057–2061. IEEE.

Wang, Z.; Fang, Z.; Wang, J.; and Yang, Y. 2020b. Vitaa: Visual-textual attributes alignment in person search by natural language. In *Proc. of ECCV*, 402–420. Springer.

Wang, Z.; Zhu, A.; Xue, J.; Jiang, D.; Liu, C.; Li, Y.; and Hu, F. 2022a. SUM: Serialized Updating and Matching for text-based person retrieval. *Knowledge-Based Systems*, 248: 108891.

Wang, Z.; Zhu, A.; Xue, J.; Wan, X.; Liu, C.; Wang, T.; and Li, Y. 2022b. Caibc: Capturing all-round information beyond color for text-based person retrieval. In *Proc. of ACM MM*, 5314–5322.

Yan, S.; Dong, N.; Zhang, L.; and Tang, J. 2022. CLIP-Driven Fine-grained Text-Image Person Re-identification. *arXiv preprint arXiv:2210.10276*.

Yao, H.; Zhang, S.; Hong, R.; Zhang, Y.; Xu, C.; and Tian, Q. 2019. Deep representation learning with part loss for person re-identification. *IEEE Transactions on Image Processing*, 28(6): 2860–2871.

Yao, L.; Huang, R.; Hou, L.; Lu, G.; Niu, M.; Xu, H.; Liang, X.; Li, Z.; Jiang, X.; and Xu, C. 2021. FILIP: Fine-grained Interactive Language-Image Pre-Training. In *Proc. of ICLR*.

Ye, M.; Shen, J.; Lin, G.; Xiang, T.; Shao, L.; and Hoi, S. C. 2021. Deep learning for person re-identification: A survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6): 2872–2893.

Zeng, H.; Zhou, T.; Wu, X.; and Cai, Z. 2022. Never Too Late: Tracing and Mitigating Backdoor Attacks in Federated Learning. In *Proc. of SRDS*, 69–81.

Zhang, Y.; and Lu, H. 2018. Deep cross-modal projection learning for image-text matching. In *Proc. of ECCV*, 686–701.

Zhao, S.; Zhu, L.; Wang, X.; and Yang, Y. 2022. Center-CLIP: Token Clustering for Efficient Text-Video Retrieval. In *Proc. of SIGIR*.

Zheng, K.; Lan, C.; Zeng, W.; Liu, J.; Zhang, Z.; and Zha, Z.-J. 2021. Pose-guided feature learning with knowledge distillation for occluded person re-identification. In *Proc. of ACM MM*, 4537–4545.

Zheng, K.; Liu, W.; Liu, J.; Zha, Z.-J.; and Mei, T. 2020a. Hierarchical gumbel attention network for text-based person search. In *Proc. of ACM MM*, 3441–3449.

Zheng, Z.; Zheng, L.; Garrett, M.; Yang, Y.; Xu, M.; and Shen, Y.-D. 2020b. Dual-path convolutional image-text embeddings with instance loss. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 16(2): 1–23.

Zhong, Z.; Zheng, L.; Kang, G.; Li, S.; and Yang, Y. 2020. Random erasing data augmentation. In *Proc. of AAAI*, volume 34, 13001–13008.

Zhou, T.; Cai, Z.; Liu, F.; and Su, J. 2023. In Pursuit of Beauty: Aesthetic-Aware and Context-Adaptive Photo Selection in Crowdsensing. *IEEE Transactions on Knowledge and Data Engineering*.

Zhu, A.; Wang, Z.; Li, Y.; Wan, X.; Jin, J.; Wang, T.; Hu, F.; and Hua, G. 2021. DSSL: Deep Surroundings-person Separation Learning for Text-based Person Retrieval. In *Proc. of ACM MM*, 209–217.

Zhu, K.; Guo, H.; Liu, Z.; Tang, M.; and Wang, J. 2020. Identity-guided human semantic parsing for person re-identification. In *Proc. of ECCV*, 346–363. Springer.

Zhuo, J.; Chen, Z.; Lai, J.; and Wang, G. 2018. Occluded person re-identification. In *Proc. of ICME*, 1–6. IEEE.