UFineBench: Towards Text-based Person Retrieval with Ultra-fine Granularity

Jialong Zuo[1,2*]   Hanyu Zhou[1]   Ying Nie[2]   Feng Zhang[1]   Tianyu Guo[2]   Nong Sang[1]
Yunhe Wang[2†]   Changxin Gao[1†]

[1]National Key Laboratory of Multispectral Information Intelligent Processing Technology,
School of Artificial Intelligence and Automation, Huazhong University of Science and Technology
[2]Huawei Noah's Ark Lab

Abstract

*Existing text-based person retrieval datasets often have relatively coarse-grained text annotations. This hinders the model to comprehend the fine-grained semantics of query texts in real scenarios. To address this problem, we contribute a new benchmark named **UFineBench** for text-based person retrieval with ultra-fine granularity.*

*Firstly, we construct a new **dataset** named UFine6926. We collect a large number of person images and manually annotate each image with two detailed textual descriptions, averaging 80.8 words each. The average word count is three to four times that of the previous datasets. In addition of standard in-domain evaluation, we also propose a special **evaluation paradigm** more representative of real scenarios. It contains a new evaluation set with cross domains, cross textual granularity and cross textual styles, named UFine3C, and a new evaluation metric for accurately measuring retrieval ability, named mean Similarity Distribution (mSD). Moreover, we propose CFAM, a more efficient **algorithm** especially designed for text-based person retrieval with ultra fine-grained texts. It achieves fine granularity mining by adopting a shared cross-modal granularity decoder and hard negative match mechanism.*

With standard in-domain evaluation, CFAM establishes competitive performance across various datasets, especially on our ultra fine-grained UFine6926. Furthermore, by evaluating on UFine3C, we demonstrate that training on our UFine6926 significantly improves generalization to real scenarios compared with other coarse-grained datasets. The dataset and code will be made publicly available at https://github.com/Zplusdragon/UFineBench.

1. Introduction

Existing text-based person retrieval benchmarks [6, 17, 54], even if claimed to be fine-grained, often have coarse-grained text annotations in practice. This makes them degenerated into attribute-based retrieval [5, 8, 31, 40] due to the provided coarse-grained descriptions to some extent. Considering this, we propose a benchmark named *UFineBench* for text-based person retrieval with ultra-fine granularity, which is more in line with real scenarios.

Our work is motivated by three main aspects. As the first aspect, existing datasets [6, 17, 54] suffer from a common issue that the text is not fine-grained enough to effectively apply to real scenarios. Specifically, as shown in Figure 1 (a), they almost only briefly describe the common appearance of persons, and lack further specific descriptions of the unique appearance. This can easily lead to the model only being able to identify typical attribute characteristics and cannot understand the fine-grained semantics of complex query texts in real scenarios. Meanwhile, as shown in Figure 1 (b), they suffer from the ambiguity of one identity-binding text corresponding to multiple different identities, hindering the model from accurately understanding how texts and images match during training. The detailed explanations can be found in Section 3.1.

As the second aspect, existing standard evaluation sets [6, 17, 54] all have fixed domain, fixed textual granularity and fixed textual styles. However, in real scenarios, there are usually three common features. 1) Extensive time and location coverage of surveillance videos, leading to substantial domain variations within the image gallery; 2) Inconsistency of granularity within the query texts, resulting from the variability in the actual information available for the person being searched; 3) The language expression of each describer has an unique style, even when conveying the same meaning. However, existing evaluation sets with fixed settings are inadequate for effectively assessing the model's performance in real scenarios with these features.

As the third aspect, existing evaluation metrics [12, 17] are not accurate enough to measure the retrieval ability. Given a text query, all images in the gallery are ranked according to their similarities with the query. The commonly used rank-k metric is calculated according to whether any image of the corresponding person is retrieved among the top k images. However, this calculation method of discretiz-

Left: A woman wearing a black Jacket and Leggings holding a bag in her left hand.

Right: She has shoulder length black hair. She is also wearing glasses and is wearing all black clothes .

(a) Coarse-grained

Left: The young man is wearing a black shirt and black jeans. He has on white shoes.

Right: The man walking and looking down at his cellphone is dressed all in black except for his white shoes.

(b) Ambiguity

This is a young man with a build that looks muscular. His skin is fair, and he has short black hair. He is wearing a blue baseball cap on his head and black-framed glasses on his face. His upper body is dressed in a white pure cotton short-sleeved T-shirt, which just reaches the position below his waist. His lower body is wearing a long grayish-milky casual pants, which just reaches the calf position. Meanwhile, there is something yellow in his right back pants pocket. On his feet are black sports shoes with white soles. He is carrying a brown crossbody bag on his back and a white plastic shopping bag with something red in his hand.

This is a mature male with a tall and sturdy physique. His exposed skin is white and smooth. He has short, white hair that is roughly at ear level. He is wearing a white short-sleeved shirt on his upper body. He is holding a red bag in his left hand, and a black briefcase is slung diagonally over his right shoulder. A black leather belt is worn around his waist. His lower body is wearing a pair of long, light khaki-colored casual pants, with the length reaching the heel of his feet. The man is wearing a pair of black casual shoes on his feet.
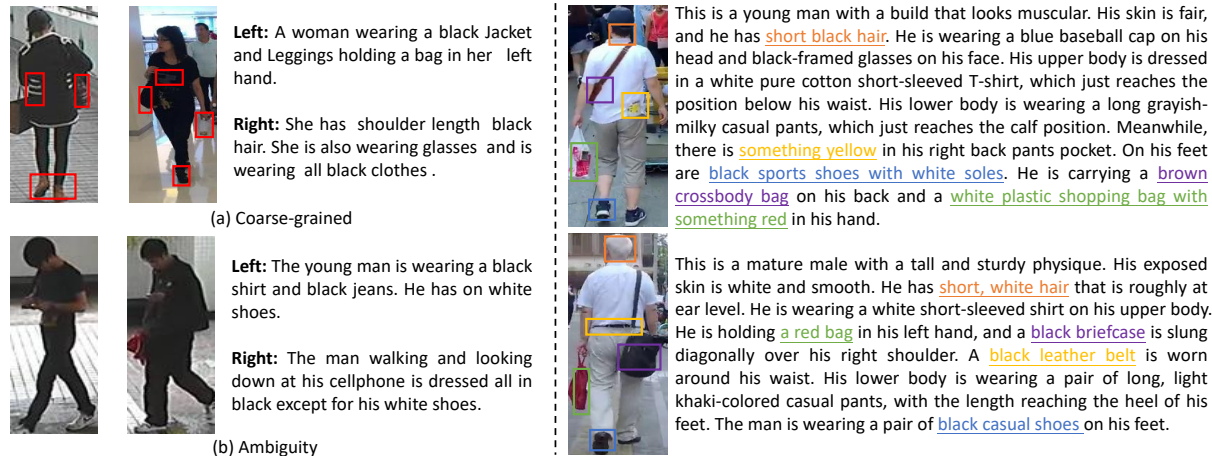
Figure 1. Comparisons between our proposed UFine6926 and existing other datasets. (a)-(b) are the examples from CUHK-PEDES [17]. In (a), some fine-grained features not described in the text are highlighted in red boxes. In (b), the text does not provide enough details to closely match its intended identity but effectively describes the other identity. Meanwhile, two examples from UFine6926 are presented on the right, with ultra fine-grained texts. As the text details some fine-grained features in the images (highlighted in different colors correspondingly), it not only provides rich cross-modal information but also effectively distinguishes highly similar image samples.

ing continuous similarity values leads to inaccurate measurement. For example, for the same rank conditions, the similarity conditions are highly likely to be different, but the rank-k metric cannot measure such differences.

Considering the above three aspects, this paper makes the following contributions. The first contribution is the build of a high quality dataset with ultra-fine granularity for text-based person retrieval, named *UFine6926*. It contains 6,926 identities, 26,206 images and 52,412 textual descriptions. A total of 58 annotators participated in crafting the textual descriptions. Each annotator is required to provide a description according to the person's appearance as detailed as possible. Compared to existing datasets [6, 17, 54], the UFine6926 dataset has significant superiority in terms of textual granularity. As shown in Figure 1, the level of detail in each description has greatly improved. The average word count is 80.8 and 3 to 4 times that of the previous datasets.

The second contribution is the construction of a special evaluation set with cross domains, cross textual granularity and cross textual styles, named *UFine3C*, which is more representative of real scenarios. It is collected from the test sets of the coarse-grained CUHK-PEDES [17], the medium-grained ICFG-PEDES [6] and our fine-grained UFine6926 to contain different domains, textual granularity and textual styles. Meanwhile, we utilize the large language models Qwen-14B [1] and Llama2-70B [37] to further enrich the variations of textual granularity and styles. It contains 7,446 images to be searched and 37,939 text queries of 2,250 persons in total.

As the third contribution, a more accurate evaluation metric is proposed for measuring retrieval ability, named mean Similarity Distribution (mSD). It is based on the continuous similarity values rather than the discrete rank con-

ditions [12, 17, 50, 52]. It requires the model to distinguish as much as possible the similarity differences between text queries and positive-negative image samples within a more precise numerical range. For the same rank conditions with different similarity conditions, it can sensitively measure the differences among them, while other metrics cannot.

Based on the proposed cross-modal shared granularity decoder and hard negative match mechanism, we also contribute a novel Cross-modal Fine-grained Aligning and Matching framework (CFAM). It establishes competitive performance on various datasets without bells and whistles, especially on our fine-grained UFine6926.

2. Related Work

CUHK-PEDES [17] is the first benchmark focusing on text-based person retrieval. It contains 40,206 images and 80,412 texts of 13,003 identities. The average word count per text is 23.5. To provide a baseline algorithm, the authors propose GNA-RNN which introduces the gated neural attention mechanism into an recurrent neural network.

However, the texts in CUHK-PEDES contain identity-irrelevant details. To address this issue, the ICFG-PEDES benchmark [6] is constructed. There are 54,522 person images, 54,522 texts of 4,102 identities gathered from MSMT17 [44]. The average word count per text is is 37.2. The authors propose SSAN to implement semantically self-alignment and part-level feature automatic extraction.

Meanwhile, RSTPReid [54] is constructed. It contains 20,505 images and 51,010 textual descriptions of 4,101 persons totally. The average word count per text is is 26.5. As a baseline algorithm, the authors propose DSSL which takes surroundings-person separation, fusion mechanism and five alignment paradigms into a unified framework.

3. Benchmark

3.1. Granularity Matters

Granularity-related research has become a hot topic in the computer vision field [11, 14–16, 18, 19, 25, 39, 47]. However, when directing to text-based person retrieval, researchers often confine themselves to a few coarse-grained benchmarks [6, 17, 54], thereby overlooking the significance of granularity in practical applications. We believe that the coarseness of textual granularity in existing benchmarks can give rise to the following two issues.

On the one hand, a substantial amount of coarse-grained descriptions greatly degrade the task into attribute-based retrieval [9, 29, 33, 34]. A simple example is illustrated in Figure 1 (a). Since the text does not describe such fine-grained features highlighted in red boxes, the model can not understand what brown boots, white stripes, and so on, refer to. When facing real scenarios with highly detailed text queries, the models trained on such coarse-grained data often prove inadequate. Meanwhile, searching for images based on coarse-grained attributes is what attribute-based person retrieval excel at. Therefore, coarse textual granularity makes these two tasks being fundamentally equivalent.

On the other hand, coarse textual granularity introduces significant ambiguity into the training process and undermines the model's performance. As a standard practice, the optimization objectives are based on the premise that each text is only associated with the images of one identity. However, in the existing benchmarks [6, 17, 54], it is common that one text can be used to describe the images from different identities. A simple example is illustrated in Figure 1 (b). Due to the overly coarseness, the text of each images cannot be highly correlated to its respective identity. Instead, it describes the other identity quite well. This ambiguity significantly hinder the model from accurately understanding how texts and images match during training.

Consequently, we emphasize that the text granularity matters and is a non-ignorable factor for text-based person retrieval. Motivated by this, we propose this benchmark with ultra-fine granularity in textual descriptions.

3.2. Dataset with Ultra-fine Granularity

We construct the first high quality dataset with ultra-fine granularity for this task, named **UFine6926**. It contains 26,206 images and 52,412 descriptions of 6,926 persons totally. The construction process is described as two steps:

First, while the person images in existing datasets are mostly derived from fixed-scene videos captured by stationary cameras, our dataset leverages a vast collection of unrestricted scene videos from the internet to obtain these images. We utilize the FairMOT algorithm [49] to extract person tracklets from the scene videos provided by [7]. One person tracklet is considered as one identity. Then, we uti-

Dataset	Maximum	Minimum	Average	Unique
CUHK-PEDES	96	15	23.5	9408
ICFG-PEDES	83	9	37.2	5790
RSTPReid	70	11	26.5	3138
UFine6926	**218**	**30**	**80.8**	**8475**

Table 1. Some statistics of texts in existing datasets. The text granularity of ours far exceeds that of others.

lize the noise-filtering strategies proposed in PLIP [55] to perform preliminary denoising on the obtained images. Finally, we conduct meticulous manual selection to ensure the image quality. Through this procedure, we have collected 26,206 high quality images of 6,926 identities in total.

Second, to obtain the ultra fine-grained textual descriptions, we hire 58 unique workers involved in the annotation task, instructing them to describe all important characteristics in the given images as detailed as possible. There are a total of 8,475 unique words in our dataset. Each person image is annotated with two textual descriptions. The longest description has 218 words and the average word count is 80.8, which is significantly larger than the 23.5 words of CUHK-PEDES [17], 37.2 words of ICFG-PEDES [6] and 26.5 words of RSTPReid [54]. As demonstrated by the examples in Figure 1 and the specific statistics provided in Table 1, our dataset exhibits a significant advantage in terms of textual granularity when compared to existing datasets.

In conclusion, the properties of our UFine6926 dataset can be summarized as follows: ultra fine-grained and unfixed scene. It can be served as a benchmark to facilitate further development in this research field.

3.3. Evaluation Set with Cross Settings

To better evaluate the model performance in real scenarios, we construct a evaluation set named **UFine3C** with cross domains, cross textual granularity and cross textual styles based on two existing datasets [6, 17] and our UFine6926. This evaluation set is very challenging and the construction process is described as two steps:

First, we collect the images and textual descriptions of according persons from the test sets of the coarse-grained "CUHK-PEDES" [17], the medium-grained "ICFG-PEDES" [6] and our fine-grained "UFine6926". We collect 750 persons from each of them to avoid bias and ensure fairness. After this collection, we obtain a set with spanning domains, textual granularity and textual styles.

Second, as large language models [2, 4, 22–24, 27, 35, 36, 51] show remarkable ability in natural language processing, we utilize Qwen-14B [1] and Llama2-70B [37] to further enrich the variations of textual granularity and styles. Given an original description, we ask the models to response with the prompt instruction: "*Please reorganize the description in a different way. You can write it as long or as short as you like*: [*original description*]". Meanwhile, we

manually revise the responses generated by them to avoid incorrect answers. Through this approach, we obtain more textual descriptions with different styles and granularity.

UFine3C contains 7,446 images, 37,939 text queries of 2,250 persons totally. This evaluation set with cross settings is more consistent with real scenarios and can be served as a standard evaluation set to facilitate relevant researches.

3.4. A New Evaluation Metric

Current benchmarks [6, 17, 54] typically use the mean average precision (mAP) to evaluate the overall performance of person retrieval algorithms. This evaluation metric is based on discrete rank conditions and cannot sensitively measure the differences in model performance at a continuous similarity level. However, continuous similarity values more realistically reflect the model's retrieval ability. As seen in Figure 2, there is a significant difference in the actual similarity values of these three rank lists. However, the APs of them both equal to 0.833, which fail to provide a fair comparison of the quality between these three rank lists.
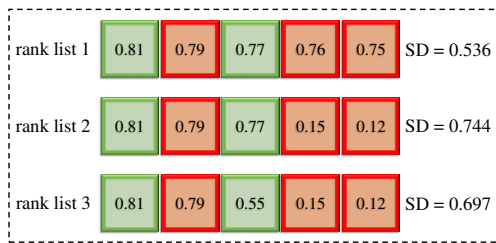

Figure 2. A toy example of the difference between SD and AP metrics. Green and red boxes mean true and false matches, respectively. For these three rank lists, the AP remains 0.833. But SD = 0.536, 0.744 and 0.697, respectively.

For UFine6926 dataset, the fine-grained retrieval ability is what we especially emphasize and any difference is a reflection of it. Therefore, we propose a new metric named mean similarity distribution (mSD) to evaluate the overall performance at a continuous similarity level. As shown in Figure 2, when mSD is used, the differences between these three rank lists can be well distinguished. The SDs of them are 0.536, 0.744 and 0.697, respectively.

Given a rank list $\{s_i\}_{i=1}^{n}$ with n ranked samples, where s_i means the similarity value of the i-th ranked sample, which is linearly normalized to the range of 0 to 1, and s^+ and s^- means respective matched and unmatched samples. The calculation process of this metric is as follows:

First, we calculate the normalized average similarity ratio between matched samples and unmatched samples by:

$$PNR = 1 - e^{-kx}, \qquad (1)$$

where x is the average similarity ratio between matched and unmatched samples in a list and k is set to 1 as default.

Then, we calculate the average similarity precision by:

$$ASP = \frac{1}{n^+} \sum_{k=1}^{n^+} \frac{\sum_{i=1}^{j_k} s_i^+}{\sum_{i=1}^{j_k} s_i}, \qquad (2)$$

where $\{j_k\}_{k=1}^{n^+}$ means the rankings of n^+ matched samples.

Then, the similarity distribution (SD) of a rank list can be measured by the product of PNR and ASP. Finally, the mean value of SDs of all rank lists, *i.e.*, mSD, is calculated as our evaluation metric.

3.5. Evaluation Paradigm

During evaluation, all images in the gallery are ranked according to their similarities with the text query. We adopt the traditional rank-k accuracy and mAP, and our newly proposed mSD to evaluate the retrieval performance.

Standard Evaluation. As a standard in-domain evaluation paradigm, UFine6926 is divided into two subsets for training and test. The training set contains 18,577 images and 37,154 texts of 4926 identities. The test set contains 7,629 images and 15,258 text queries of 2,000 identities.

Special Evaluation. As a special evaluation paradigm with cross settings, UFine3C is utilized as a test set for evaluating the model performance in real scenarios. The training set is the same as that of standard paradigm.

4. Method

4.1. Overview

In this section, we introduce a Cross-modal Fine-grained Aligning and Matching framework (CFAM), which achieves fine granularity mining in a non trivial way. The whole framework is shown in Figure 3, given an input image I and an input text T, the CLIP [26] pre-trained visual encoder $\mathbf{E_v}$ and textual encoder $\mathbf{E_t}$ are adopted to extract the visual embeddings $\mathbf{V} = \{\mathbf{v_1}, \mathbf{v_2}, \ldots, \mathbf{v_{n_i}}\}$ and textual embeddings $\mathbf{W} = \{\mathbf{w_1}, \mathbf{w_2}, \ldots, \mathbf{w_{n_t}}\}$, respectively. Specially, we design a cross-modal fine-grained align and match module to improve the fine-grained retrieval ability. Through a shared cross-modal granularity decoder and hard negative match mechanism, the framework achieves competitive performance on various datasets and settings.

4.2. Cross-modal Fine-grained Aligning

Given the extracted visual and textual embeddings, most existing methods [12, 55] only calculate global similarities to achieve cross-modal alignment, which is inclined to overlooking the fine-grained details in both modalities. Therefore, we propose to perform more fine-grained alignment based on the local embeddings. However, the visual embeddings $\mathbf{V}$ and textual embeddings $\mathbf{W}$ usually have different length. To address this issue, we propose a shared cross-modal granularity decoder $\mathbf{D_g}$ with a fixed set of granularity
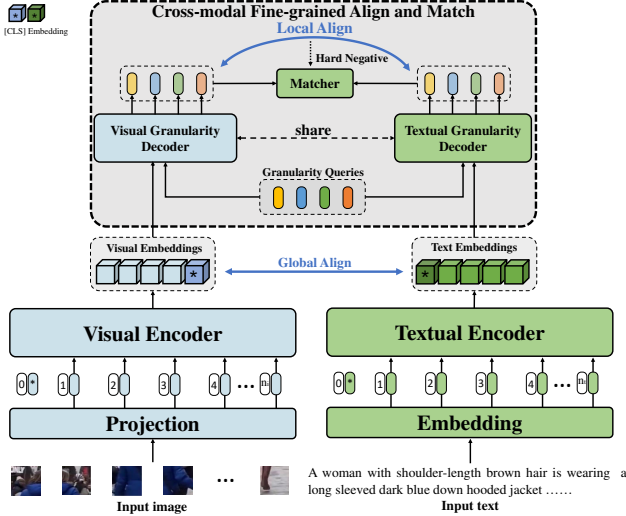
Figure 3. Overview of the proposed CFAM framwork.

queries $\mathbf{Q} = \{\mathbf{q_1}, \mathbf{q_2}, \dots, \mathbf{q_K}\}$. These queries can interact with the embeddings and extract fine-grained information for cross-modal alignment.

For visual fine-grained information extraction, the granularity decoder $\mathbf{D_g}$ take the queries $\mathbf{Q}$ and the visual embeddings $\mathbf{V}$ as input, and then produce the fine-grained visual representations as follow,

$$\overline{\mathbf{V}} = \mathbf{D_g}(\mathbf{Q}, \mathbf{V}), \tag{3}$$

where $\overline{\mathbf{V}} = \{\overline{\mathbf{v}}_1, \overline{\mathbf{v}}_2, \dots, \overline{\mathbf{v}}_K\}$ has the same length as the granularity queries. Meanwhile, the fine-grained textual representations $\overline{\mathbf{W}} = \{\overline{\mathbf{w}}_1, \overline{\mathbf{w}}_2, \dots, \overline{\mathbf{w}}_K\}$ are produced in the similar way.

In this decoding procedure, the output representations corresponding to a certain query contain the relevant fine-grained information from both modalities and share similar semantic content. Therefore, the cross-modal similarity for each query output can be measured to achieve fine-grained alignment. We use the cosine distance to measure the similarity and the overall similarity of the K query outputs can be calculated by:

$$Sim(\mathbf{V}, \mathbf{W}) = \frac{1}{K} \sum_{i}^{K} \frac{\overline{\mathbf{v}}_i^{\mathsf{T}} \overline{\mathbf{w}}_i}{||\overline{\mathbf{v}}_i|| \, ||\overline{\mathbf{w}}_i||}. \tag{4}$$

Then, given a batch of B image-text pairs, the commonly used SDM loss [12] will be utilized to calculate the local alignment loss $\mathcal{L}_{ls}$ according to the similarity distribution.

4.3. Cross-modal Hard Negative Matching

To further facilitate the cross-modal alignment, we propose to perform prediction on whether the granularity representations of each modality are matched. This task can be seen as a binary classification problem: the paired image-text is considered the positive sample, while the unpaired

text is considered the negative one. Unlike common random sampling, we employ the hard negative mining strategy, which is beneficial to learning more discriminative representations.

For each image within a batch $|B|$, we sample the unpaired text whose owns the highest similarity with this image as the hard negative. Also, we sample one hard negative image for each text in the same way. Through this approach, we obtain $|B|$ positive pairs and $2|B|$ negative pairs, denoted as $|\overline{B}|$ pairs. Then, we pass the fine-grained representations of these $|\overline{B}|$ pairs through a binary classifier named Matcher, to optimize the following objective:

$$\mathcal{L}_m = \frac{1}{|\overline{B}|} \sum_{(\overline{\mathbf{V}}, \overline{\mathbf{W}}) \in \overline{B}} (\hat{y} \log p(\overline{\mathbf{V}}, \overline{\mathbf{W}}) + (1 - \hat{y})(1 - \log p(\overline{\mathbf{V}}, \overline{\mathbf{W}}))), \tag{5}$$

where p is a binary likelihood distribution function, and $\hat{y}$ is 1 if $(\overline{\mathbf{V}}, \overline{\mathbf{W}})$ is matched, 0 otherwise.

4.4. Training and Inference Strategy

As complementary to fine-grained alignment, we compute the global similarity between the global visual embedding and the global textual embedding, and optimize the global alignment loss $\mathcal{L}_{gs}$ according to it. Also, we propose to utilize the cross-modal identity classifying loss $\mathcal{L}_{cid}$ with hard negative samples to explicitly ensure that the representations of the same image/text pair are closely clustered together. The details of this strategy are shown in the supplement. The overall training objective is the weighted sum of the above losses:

$$\mathcal{L} = \mathcal{L}_{gs} + \lambda_1 \mathcal{L}_{ls} + \lambda_2 \mathcal{L}_m + \lambda_3 \mathcal{L}_{cid}, \tag{6}$$

where $\lambda_1, \lambda_2, \lambda_3$ are hyper-parameters to adjust the weight of each loss, which are all set to 1 as the default.

In the inference phase, we will discard all additional designs and only compare the similarities between the visual and textual global embeddings. First, all images in the gallery will be passed to the visual encoder to extract the according global visual embeddings. Second, for each text query, we obtain its textual global embedding in a similar way and then we compute its similarity with the visual global embeddings of all images. Finally, we utilize the calculated similarities for ranking the image candidates.

5. Experiments

5.1. Implementation

We conduct text-based person retrieval on our proposed fine-grained UFine6926 datasets and three existing datasets CUHK-PEDES [17], ICFG-PEDES [6] and RST-PReid [54]. Meanwhile, we utilize the UFine3C evaluation set as an extra supplement to assess the generalization ability of the models in real scenarios. We adopt the

popular rank-k metric (k=1,5,10), the mean Average Precision (mAP) and our proposed mean Similarity Distribution (mSD) as the evaluation metrics. The higher rank-k, mAP and mSD indicates better performance.

CFAM mainly consists of a pre-trained visual encoder, *i.e.*, CLIP-ViT-B/16 [26], a pre-trained text encoder, *i.e.*, CLIP textual encoder, a random-initialized granularity decoder and a matcher. The granularity decoder is shared by visual and textual modalities, consisting of 2-layer transformer blocks [38]. The matcher is consisted of 2-layer transformer blocks and an MLP with sigmoid activation. For each layer of the granularity decoder and matcher, the hidden size and number of heads are set to 512 and 8. The number of granularity queries is set to 16 and their hidden dimension is 512. For downstream training, the images are resized to 384×128 and the maximum length of the textual tokens is set to 168. The batchsize per GPU is set to 64. Also, random erasing, horizontally flipping and crop with padding are employed for image augmentation. Random masking and replacement is employed for text augmentation. Our CFAM is trained with Adam [13] for 60 epochs with an initial learning rate $1e^{-5}$. We adopt the linearly warm-up strategy within the beginning 5 epochs. For the random-initialized modules, the initial learning rate is set to $5e^{-5}$. We adopt the cosine learning rate decay strategy. The experiments are performed on 1 V100 32GB GPU.

5.2. Importance of Fine Granularity

In this section, we conduct experiments to study the importance of fine granularity in real-world scenarios. Specifically, we have trained two baseline models (PLIP [55] and IRRA [12]) and our CFAM on three coarse-grained existing datasets and our fine-grained dataset. Then, due to the fact that generalization ability is indispensable in real-world scenarios, we evaluate their performance under a range of cross settings. Please note that PLIP [55] is a pre-trained model on a large amount of pedestrian data, giving it a SoTA generalization capability in the current field. However, compared to PLIP, CFAM still demonstrates competitive performance under a range of cross settings.

Fineness Better Generalizes to Real-world Scenarios. In real-world scenarios, there are many variations in image domains, textual granularity and textual styles. The ability to effectively address these variations is a necessity for a high-level model. To study whether training on our proposed fine-grained UFine6926 dataset can lead the model to better generalize to the real-world scenarios than existing coarse-grained datasets [6, 17, 54], we conduct experiments by setting the UFine3C evaluation set as the target set to be transfered. The specific experimental procedure is described as follows. Firstly, We choose two existing popular open-source and state-of-the-art methods [12, 55] and our CFAM as the baseline models. Secondly, we

train the models on each training set of the three coarse-grained datasets [6, 17, 54] and our fine-grained UFine6926. Thirdly, we directly evaluate the trained models' performance on the UFine3C evaluation set. By comparing the performance differences, we can effectively assess the generalization ability of models trained on various datasets to real-world scenarios. The experimental results are reported in Table 2. As we can see, for all the three baseline models, training on our UFine6926 dataset will significantly lead to better performance on the UFine3C evaluation set. Specifically, CFAM achieves 62.84%, 77.82%, 83.23% and 46.04% on rank-1, rank-5, rank-10 and mSD, respectively, greatly exceeding the results obtained by training on other coarse-grained datasets. We must note that the training samples in UFine6926 is much less than that in other datasets, while still achieves state-of-the-art performance. The results demonstrate that our fine-grained UFine6926 helps to learn more discriminative and general representations, which is beneficial to generalize to the real scenarios.

Generalization between Fineness and Coarseness. We conduct the experiments under two aspects. As the first aspect, we investigate the differences in mutual generalization capabilities between coarse-grained and fine-grained datasets. We train the models on the coarse-grained datasets and then directly transfer them to our fine-grained dataset, and vice versa. The experimental results are reported in Table 3 (a). As we can see, for all of the three baseline models, training on the coarse-grained datasets cannot well be transferred to our fine-grained dataset. For example, when transferring to UFine6926, PLIP [55] trained on CUHK-PEDES [17] only achieves 20.52%, 33.74%, 42.69% on rank-1, rank-5 and rank-10, respectively, which falls far short of practical application requirements. However, training on our fine-grained dataset can be transferred to the coarse-grained datasets to a better extent. This demonstrate that training on our fine-grained dataset enables generalization to coarse-grained datasets, while the reverse is not true. As the second aspect, we demonstrate that even when transferring to a coarse-grained dataset, training with our fine-grained dataset is mostly superior to using other coarse-grained datasets. We choose the coarse-grained ICFG-PEDES [6] and RSTPReid [54] datasets as the target datasets. As the results reported in Table 3 (b), for all of the baselines, even though our dataset contains very little coarse-grained data, training on our UFine6926 still achieves better or competitive performance on the two target coarse-grained datasets. All the results demonstrate that our fine-grained dataset improves general representation learning for text-based person retrieval.

Qualitative Results. To make a more realistic comparison of the models' performance in real-world scenarios, we conduct a straightforward qualitative experiment. We choose our CFAM trained on the coarse-grained CUHK-

Training Sets	CFAM					IRRA [12]					PLIP [55]				
	R@1	R@5	R@10	mAP	mSD	R@1	R@5	R@10	mAP	mSD	R@1	R@5	R@10	mAP	mSD
CUHK-PEDES	53.80	71.05	78.25	50.40	38.26	50.06	67.98	75.46	47.57	36.50	40.45	57.51	65.20	38.94	30.82
ICFG-PEDES	36.79	54.64	62.93	34.21	25.47	30.57	47.61	55.87	28.38	21.24	34.32	50.52	57.94	32.59	24.88
RSTPReid	29.85	49.08	58.54	29.66	21.82	21.62	39.53	49.38	21.90	16.09	25.25	40.70	48.30	24.62	18.18
UFine6926	**62.84**	**77.82**	**83.23**	**59.31**	**46.04**	**56.34**	**72.17**	**78.47**	**54.24**	**42.92**	**64.59**	**80.16**	**85.63**	**60.43**	**47.76**

Table 2. Performance comparisons on the UFine3C evaluation dataset. The models are trained on the training sets of CUHK-PEDES, ICFG-PEDES, RSTPReid and our UFine6926, and then are directly evaluated on the UFine3C dataset. Although the training samples in UFine6926 is less than that in other datasets, training on it still achieves state-of-the-art performance. The bold results indicate the best.

Domains	CFAM					IRRA [12]					PLIP [55]				
	R@1	R@5	R@10	mAP	mSD	R@1	R@5	R@10	mAP	mSD	R@1	R@5	R@10	mAP	mSD
CUHK→ UFine	42.49	59.47	68.14	45.06	33.74	37.63	54.99	64.46	40.79	30.80	20.52	33.74	42.69	24.17	18.90
ICFG→UFine	20.65	34.66	43.05	23.09	16.65	14.99	26.85	33.92	17.02	12.29	12.13	21.88	28.73	14.98	11.20
RSTP→UFine	20.20	35.31	44.02	23.13	16.73	13.13	25.59	33.81	15.55	11.20	9.75	18.86	25.27	12.32	8.95
UFine→CUHK	48.72	70.21	78.17	44.42	33.23	41.41	62.72	71.85	39.22	29.86	56.53	77.24	84.10	51.60	39.85
UFine→ICFG	40.78	60.90	69.31	22.30	16.28	35.08	55.16	64.02	18.87	13.85	51.52	70.96	78.02	27.67	20.93
UFine→RSTP	45.10	72.35	81.45	35.40	25.40	41.30	64.25	76.00	32.04	22.93	43.85	72.10	80.60	33.88	24.97

(a) Differences in mutual generalization capabilities between coarse-grained and fine-grained datasets.

Domains	CFAM					IRRA [12]					PLIP [55]				
	R@1	R@5	R@10	mAP	mSD	R@1	R@5	R@10	mAP	mSD	R@1	R@5	R@10	mAP	mSD
CUHK → ICFG	46.21	65.18	72.65	24.77	18.00	42.42	62.07	69.64	21.80	15.94	53.81	72.56	79.34	30.20	22.76
RSTP→ICFG	38.55	55.37	63.53	24.66	18.46	32.37	49.71	57.75	20.57	15.44	51.01	69.52	76.71	32.14	23.82
UFine→ICFG	40.78	60.90	69.31	22.30	16.28	35.08	55.16	64.02	18.87	13.85	51.52	70.96	78.02	27.67	20.93
ICFG→CUHK	40.48	63.48	72.68	37.38	27.11	33.45	56.12	66.21	31.39	22.78	56.40	76.98	83.82	51.72	39.12
RSTP→CUHK	40.11	63.55	72.61	37.29	27.04	32.67	55.20	65.34	30.17	21.87	50.15	72.84	81.01	46.84	34.43
UFine→CUHK	48.72	70.21	78.17	44.42	33.23	41.41	62.72	71.85	39.22	29.86	56.53	77.24	84.10	51.60	39.85

(b) Fine-grained dataset can even better be transferred to other coarse-grained datasets than the coarse-grained dataset.

Table 3. Performance comparisons on the generalization performance between our fine-grained UFine6926 and three existing datasets CUHK-PEDES [17], ICFG-PEDES [6] and RSTPReid [54]. The arrow direction indicates the source dataset and the target dataset.

PEDES [17] and the fine-grained UFine6926 as the baseline models to be compared. Then, we manually provide any textual descriptions to search the according persons in the UFine3C dataset. The rank-10 retrieval results from the CFAM models trained on CUHK-PEDES and UFine6926 respectively are compared in Figure 4. As it shows, training on UFine6926 achieves more accurate retrieval results and can fully perceive fine-grained discriminative clues to distinguish different persons, while training on CUHK-PEDES fails to do so. This is illustrated in the orange highlighted text and image region box in Figure 4.

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

In this section, we compare the performance of our proposed CFAM framework with state-of-the-art (SoTA) methods on our fine-grained UFine6926 dataset and three public coarse-grained datasets [6, 17, 54].

Performance Comparisons on UFine6926. We utilize two evaluation sets for the performance comparison on UFine6926. The first is the UFine3C evaluation set. The second is the UFine6926 test set. We evaluate the performance of existing SoTA methods trained on our new UFine6926. As the results shown in Table 4, on each evaluation set, CFAM outperforms all other SoTA methods.

A man wearing a white short-sleeved shirt and black shorts has a black watch on his wrist.

The woman is wearing a black clothing and a pair of black pants, and she is wearing a pair of black glasses.

Figure 4. Comparison of rank-10 retrieval results on UFine3C between CFAM trained on UFine6926 [17] (the first row) and CUHK-PEDES (the second raw) for each textual description. The images that fully match the text are marked in green, and the unmatched ones are marked in red.

Specifically, with CLIP-ViT-L/14 [26] setting, it achieves 62.84% rank-1 accuracy, 59.31% mAP and 46.04% mSD on UFine3C, respectively. Meanwhile, it achieves 88.51% rank-1 accuracy, 57.09% mAP and 68.45% mSD on UFine6926 test set, respectively. These results demonstrate the superior fine-grained retrieval capability of CFAM.

	Methods	Metrics				
		R@1	R@5	R@10	mAP	mSD
UFine3C	NAFS [10]	43.69	61.34	69.72	39.31	30.32
	LGUR [30]	51.26	69.67	75.32	49.22	38.13
	SSAN [6]	53.67	71.15	77.15	51.40	39.66
	IRRA [12]	56.34	72.17	78.47	54.24	42.92
	CFAM(B/16)	59.61	75.82	82.90	56.53	45.22
	CFAM(L/14)	**64.62**	**79.56**	**85.13**	**61.13**	**48.34**
UFine6926	NAFS [10]	64.11	80.32	85.05	63.47	49.61
	LGUR [30]	70.69	84.57	89.91	68.93	56.23
	SSAN [6]	75.09	88.63	92.84	73.14	59.41
	IRRA [12]	83.53	92.94	95.95	82.79	66.35
	CFAM(B/16)	86.65	95.71	98.12	85.23	67.49
	CFAM(L/14)	**89.61**	**96.78**	**98.89**	**88.19**	**69.45**

Table 4. We train some state-of-the-art open-source models on UFine6926 and evaluate the performance under two evaluation settings. We show the best score in bold .

Method	CUHK-PEDES				
	R@1	R@5	R@10	mAP	mSD
MIA [20]	53.10	75.00	82.90	-	-
TIMAM [28]	54.51	77.56	79.27	-	-
TDE [21]	55.25	77.46	84.56	-	-
NAFS [10]	59.94	79.86	86.70	54.07	-
SSAN [6]	61.37	80.15	86.73	-	-
LapsCore [45]	63.40	-	87.80	-	-
TIPCB [3]	64.26	83.19	89.10	-	-
CAIBC [42]	64.43	82.87	88.37	-	-
LGUR [30]	65.25	83.12	89.00	-	-
IVT [32]	65.59	83.11	89.21	-	-
PLIP [55]	69.23	85.84	91.16	-	-
CFine [46]	69.57	85.93	91.15	-	-
IRRA [12]	73.38	89.93	93.71	66.13	51.49
CFAM(B/16)	73.67	89.71	93.57	65.94	51.32
CFAM(L/14)	**76.71**	**91.83**	**95.96**	**68.47**	**52.93**

Table 5. Comparison with the state-of-the-art methods on CUHK-PEDES [17]. We show the best score in bold.

Performance Comparisons on Other Datasets. The experimental results on the CUHK-PEDES [17], ICFG-PEDES [6] and RSTPReid [54] datasets are reported in Table 5, Table 6 and Table 7, respectively. On CUHK-PEDES, with the CLIP-ViT-B/16 setting, CFAM achieves competitive results to recent state-of-the-art methods without bells and whistles, achieving 72.87% rank-1 accuracy, 64.92% mAP and 50.20% mSD, respectively. Meanwhile, with the CLIP-ViT-L/14 setting, the performance of CFAM can be further improved, achieving 75.60% rank-1, 67.27% mAP and 51.83% mSD, respectively. This means that our method has good scalability. On ICFG-PEDES and RSTPReid, our CFAM also outperforms all state-of-the-art methods by a considerable margin. It achieves 65.38% rank-1, 39.42% mAP and 30.29% mSD on ICFG-PEDES, 62.45% rank-1, 49.50% mAP and 36.92% mSD on RSTPReid, respectively. The results demonstrate that CFAM helps to learn general representations on various datasets.

5.4. Ablation Study

To verify the contribution of each component in CFAM, we conduct an ablation experiment on CUHK-PEDES dataset [17]. The results are reported in Table 8. No.0 is the baseline utilizing the original InfoNCE loss in CLIP [26]

Method	ICFG-PEDES				
	R@1	R@5	R@10	mAP	mSD
Dual Path [53]	38.99	59.44	68.41	-	-
CMPM/C [48]	43.51	65.44	74.26	-	-
ViTAA [41]	50.98	68.79	75.78	-	-
SSAN [6]	54.23	72.63	79.53	-	-
LGUR [30]	59.02	75.32	81.56	-	-
IVT [32]	56.04	73.60	80.22	-	-
CFine [46]	60.83	76.55	82.42	-	-
IRRA [12]	63.46	80.25	85.82	38.06	29.54
PLIP [55]	64.25	80.88	86.32	-	-
CFAM(B/16)	63.57	80.57	86.32	38.34	29.01
CFAM(L/14)	**66.58**	**82.47**	**87.37**	**40.46**	**31.78**

Table 6. Comparison with the state-of-the-art methods on ICFG-PEDES [6]. We show the best score in bold.

Method	RSTPReid				
	R@1	R@5	R@10	mAP	mSD
DSSL [54]	39.05	59.44	68.41	-	-
SSAN [6]	43.50	67.80	77.15	-	-
LBUL [43]	45.55	68.20	77.85	-	-
IVT [32]	46.70	70.00	78.80	-	-
CFine [46]	50.55	72.50	81.60	-	-
IRRA [12]	60.20	81.30	88.20	47.17	35.22
CFAM(B/16)	60.51	82.85	89.71	47.64	35.67
CFAM(L/14)	**63.54**	**84.75**	**92.32**	**50.48**	**37.96**

Table 7. Comparison with the state-of-the-art methods on RST-PReid [54]. We show the best score in bold.

No.	Components				CUHK-PEDES				
	$\mathcal{L}_{gs}$	$\mathcal{L}_{ls}$	$\mathcal{L}_{m}$	$\mathcal{L}_{vap}$	R@1	R@5	R@10	mAP	mSD
0					68.45	86.50	91.68	61.28	46.31
1	✓				70.42	87.20	92.22	63.00	48.38
2	✓	✓			71.69	87.87	92.37	63.85	49.12
3	✓	✓	✓		72.42	88.31	92.80	64.81	49.84
4	✓	✓	✓	✓	**73.67**	**89.71**	**93.57**	**65.94**	**51.32**

Table 8. Ablation study on each component of CFAM.

to align the cross-modal global embeddings. As we can see, each component facilitates the model's capability, and combining all of them leads to the best performance. In addition, we have conducted further ablation experiments, which are detailed in the supplement.

6. Conclusion

This paper introduces a new benchmark for text-based person retrieval with ultra-fine granularity. We firstly contribute a manually annotated dataset named UFine6926 with ultra fine-grained texts. Meanwhile, we propose a special evaluation paradigm more representative for real scenarios with a new evaluation set named UFine3C and a new metric named mSD. Then, CFAM is proposed in the attempt to achieve fine-grained cross-modal representation correlation. Our benchmark will enable research possibilities in multiple directions, *e.g.*, fine-grained retrieval, real scenario generalization, multi-granularity adaptation, efficient structure, *etc*. We believe this work will shed light on more future researches in this community.

References

[1] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023. 2, 3

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *NeurIPS*, 33:1877–1901, 2020. 3

[3] Yuhao Chen, Guoqing Zhang, Yujiang Lu, Zhenxing Wang, and Yuhui Zheng. Tipcb: A simple but effective part-based convolutional baseline for text-based person search. *Neurocomputing*, 494:171–181, 2022. 8

[4] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022. 3

[5] Mickael Cormier, Andreas Specker, Julio Junior, CS Jacques, Lucas Florin, Jürgen Metzler, Thomas B Moeslund, Kamal Nasrollahi, Sergio Escalera, and Jürgen Beyerer. Upar challenge: Pedestrian attribute recognition and attribute-based person retrieval–dataset, design, and results. In *WACV*, pages 166–175, 2023. 1

[6] Zefeng Ding, Changxing Ding, Zhiyin Shao, and Dacheng Tao. Semantically self-aligned network for text-to-image part-aware person re-identification. *arXiv preprint arXiv:2107.12666*, 2021. 1, 2, 3, 4, 5, 6, 7, 8

[7] Dengpan Fu, Dongdong Chen, Hao Yang, Jianmin Bao, Lu Yuan, Lei Zhang, Houqiang Li, Fang Wen, and Dong Chen. Large-scale pre-training for person re-identification with noisy labels. In *CVPR*, pages 2476–2486, 2022. 3

[8] Hiren Galiyawala, Mehul S Raval, and Dhyey Savaliya. Dsapr: discrete soft biometric attribute-based person retrieval in surveillance videos. In *AVSS*, pages 1–7. IEEE, 2021. 1

[9] Hiren Galiyawala, Mehul S Raval, and Meet Patel. Person retrieval in surveillance videos using attribute recognition. *Journal of Ambient Intelligence and Humanized Computing*, pages 1–13, 2022. 3

[10] Chenyang Gao, Guanyu Cai, Xinyang Jiang, Feng Zheng, Jun Zhang, Yifei Gong, Pai Peng, Xiaowei Guo, and Xing Sun. Contextual non-local alignment over full-scale representation for text-based person search. *arXiv preprint arXiv:2101.03036*, 2021. 8

[11] Wei He, Kai Han, Ying Nie, Chengcheng Wang, and Yunhe Wang. Species196: A one-million semi-supervised dataset for fine-grained species recognition. *NeurIPS*, 36, 2024. 3

[12] Ding Jiang and Mang Ye. Cross-modal implicit relation reasoning and aligning for text-to-image person retrieval. In *CVPR*, 2023. 1, 2, 4, 5, 6, 7, 8

[13] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[14] Juncheng Li, Xin He, Longhui Wei, Long Qian, Linchao Zhu, Lingxi Xie, Yueting Zhuang, Qi Tian, and Siliang Tang. Fine-grained semantically aligned vision-language pre-training. *NeurIPS*, 35:7290–7303, 2022. 3

[15] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, pages 12888–12900. PMLR, 2022.

[16] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023. 3

[17] Shuang Li, Tong Xiao, Hongsheng Li, Bolei Zhou, Dayu Yue, and Xiaogang Wang. Person search with natural language description. In *CVPR*, pages 1970–1979, 2017. 1, 2, 3, 4, 5, 6, 7, 8

[18] Zhishan Li, Ying Nie, Kai Han, Jianyuan Guo, Lei Xie, and Yunhe Wang. A transformer-based object detector with coarse-fine crossing representations. *NeurIPS*, 35:38733–38746, 2022. 3

[19] Ying Nie, Wei He, Kai Han, Yehui Tang, Tianyu Guo, Fanyi Du, and Yunhe Wang. Lightclip: Learning multi-level interaction for lightweight vision-language models. *arXiv preprint arXiv:2312.00674*, 2023. 3

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

[21] Kai Niu, Yan Huang, and Liang Wang. Textual dependency embedding for person search by language. In *ACM MM*, pages 4032–4040, 2020. 8

[22] OpenAI. Chatgpt. https://openai.com/blog/chatgpt/, 2023. 3

[23] OpenAI. Gpt-4 technical report, 2023.

[24] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *NeurIPS*, 35:27730–27744, 2022. 3

[25] Zhengxin Pan, Fangyu Wu, and Bailing Zhang. Fine-grained image-text matching by cross-modal hard aligning network. In *CVPR*, pages 19275–19284, 2023. 3

[26] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021. 4, 6, 7, 8

[27] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020. 3

[28] Nikolaos Sarafianos, Xiang Xu, and Ioannis A Kakadiaris. Adversarial representation learning for text-to-image matching. In *ICCV*, pages 5814–5824, 2019. 8

[29] Arne Schumann, Andreas Specker, and Jürgen Beyerer. Attribute-based person retrieval and search in video sequences. In *AVSS*, pages 1–6. IEEE, 2018. 3

[30] Zhiyin Shao, Xinyu Zhang, Meng Fang, Zhifeng Lin, Jian Wang, and Changxing Ding. Learning granularity-unified representations for text-to-image person re-identification. In *ACM MM*, pages 5566–5574, 2022. 8

[31] Rasha Shoitan, Mona M Moussa, and Heba A El Nemr. Attribute based spatio-temporal person retrieval in video surveillance. *Alexandria Engineering Journal*, 63:441–454, 2023. 1

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

[33] Andreas Specker and Jürgen Beyerer. Improving attribute-based person retrieval by using a calibrated, weighted, and distribution-based distance metric. In *ICIP*, pages 2378–2382. IEEE, 2021. 3

[34] Andreas Specker, Mickael Cormier, and Jürgen Beyerer. Upar: Unified pedestrian attribute recognition and person retrieval. In *WACV*, pages 981–990, 2023. 3

[35] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca, 2023. 3

[36] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 3

[37] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 2, 3

[38] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 30, 2017. 6

[39] Xiaohan Wang, Linchao Zhu, Zhedong Zheng, Mingliang Xu, and Yi Yang. Align and tell: Boosting text-video retrieval with local alignment and fine-grained supervision. *IEEE TMM*, 2022. 3

[40] Yaodong Wang, Zhenfei Hu, and Zhong Ji. Attribute-wise reasoning reinforcement learning for pedestrian attribute retrieval. *International Journal of Multimedia Information Retrieval*, 12(2):35, 2023. 1

[41] Zhe Wang, Zhiyuan Fang, Jun Wang, and Yezhou Yang. Vitaa: Visual-textual attributes alignment in person search by natural language. In *ECCV*, pages 402–420. Springer, 2020. 8

[42] Zijie Wang, Aichun Zhu, Jingyi Xue, Xili Wan, Chao Liu, Tian Wang, and Yifeng Li. Caibc: Capturing all-round information beyond color for text-based person retrieval. In *ACM MM*, pages 5314–5322, 2022. 8

[43] Zijie Wang, Aichun Zhu, Jingyi Xue, Xili Wan, Chao Liu, Tian Wang, and Yifeng Li. Look before you leap: Improving text-based person retrieval by learning a consistent cross-modal common manifold. In *ACM MM*, pages 1984–1992, 2022. 8

[44] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *CVPR*, pages 79–88, 2018. 2

[45] Yushuang Wu, Zizheng Yan, Xiaoguang Han, Guanbin Li, Changqing Zou, and Shuguang Cui. Lapscore: language-guided person search via color reasoning. In *ICCV*, pages 1624–1633, 2021. 8

[46] Shuanglin Yan, Neng Dong, Liyan Zhang, and Jinhui Tang. Clip-driven fine-grained text-image person re-identification. *IEEE TIP*, 2023. 8

[47] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. Filip: Fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783*, 2021. 3

[48] Ying Zhang and Huchuan Lu. Deep cross-modal projection learning for image-text matching. In *ECCV*, pages 686–701, 2018. 8

[49] Yifu Zhang, Chunyu Wang, Xinggang Wang, Wenjun Zeng, and Wenyu Liu. Fairmot: On the fairness of detection and re-identification in multiple object tracking. *IJCV*, 129:3069–3087, 2021. 3

[50] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *ICCV*, pages 1116–1124, 2015. 2

[51] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *NeurIPS*, 36, 2024. 3

[52] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *ICCV*, pages 3774–3782, 2017. 2

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

[54] Aichun Zhu, Zijie Wang, Yifeng Li, Xili Wan, Jing Jin, Tian Wang, Fangqiang Hu, and Gang Hua. Dssl: Deep surroundings-person separation learning for text-based person retrieval. In *ACM MM*, pages 209–217, 2021. 1, 2, 3, 4, 5, 6, 7, 8

[55] Jialong Zuo, Changqian Yu, Nong Sang, and Changxin Gao. Plip: Language-image pre-training for person representation learning. *arXiv preprint arXiv:2305.08386*, 2023. 3, 4, 6, 7, 8