Instruct-ReID: A Multi-purpose Person Re-identification Task with Instructions

Weizhen He[1†] **Yiheng Deng**[1] **Shixiang Tang**[2,3*] **Qihao Chen**[4]
Qingsong Xie[5] **Yizhou Wang**[2] **Lei Bai**[2] **Feng Zhu**[3] **Rui Zhao**[3,6]
Wanli Ouyang[2] **Donglian Qi**[1] **Yunfeng Yan**[1*]
[1]Zhejiang University [2]Shanghai AI Laboratory
[3]SenseTime Research [4]Liaoning Technical University [5]Shanghai Jiao Tong University
[6]Qing Yuan Research Institute, Shanghai Jiao Tong University
hewz@zju.edu.cn, yvonnech@zju.edu.cn

Abstract

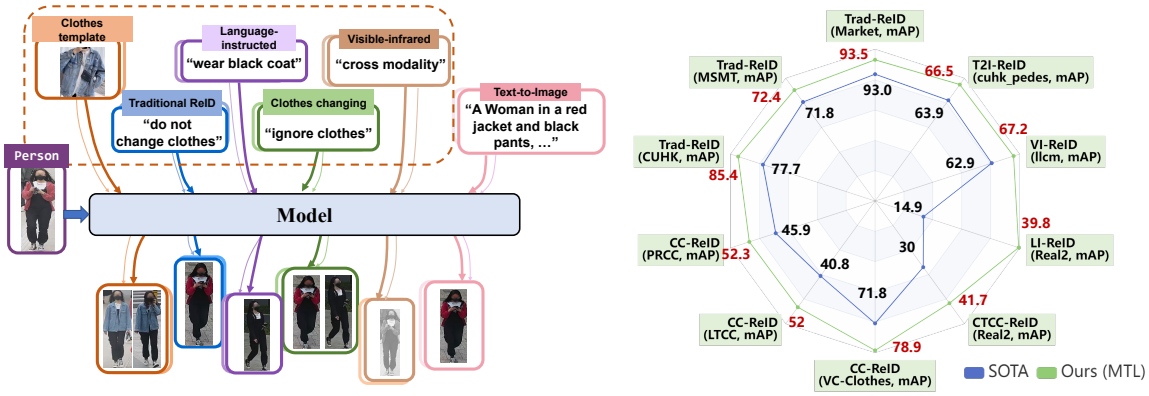
*Human intelligence can retrieve any person according to both visual and language descriptions. However, the current computer vision community studies specific person re-identification (ReID) tasks in different scenarios separately, which limits the applications in the real world. This paper strives to resolve this problem by proposing a new instruct-ReID task that requires the model to retrieve images according to the given image or language instructions. Our instruct-ReID is a more general ReID setting, where existing **6** ReID tasks can be viewed as special cases by designing different instructions. We propose a large-scale OmniReID benchmark and an adaptive triplet loss as a baseline method to facilitate research in this new setting. Experimental results show that the proposed multi-purpose ReID model, trained on our OmniReID benchmark without fine-tuning, can improve **+0.5%**, **+0.6%**, **+7.7%** mAP on Market1501, MSMT17, CUHK03 for traditional ReID, **+6.4%**, **+7.1%**, **+11.2%** mAP on PRCC, VC-Clothes, LTCC for clothes-changing ReID, **+11.7%** mAP on COCAS+ real2 for clothes template based clothes-changing ReID when using only RGB images, **+24.9%** mAP on COCAS+ real2 for our newly defined language-instructed ReID, **+4.3%** on LLCM for visible-infrared ReID, **+2.6%** on CUHK-PEDES for text-to-image ReID. The datasets, the model, and code are available at https://github.com/hwz-zju/Instruct-ReID.*

1. Introduction

Identifying individuals exhibiting significant appearance variations from multi-model descriptions is a fundamental aspect of human intelligence with broad applications [35, 37, 38, 70]. To imbue our machine with this capability, person re-identification (ReID) [8, 73, 75] has been introduced to retrieve images of the target person from a vast repository of surveillance videos or images across locations and time [11, 51]. Recently, significant advancements have been made in developing precise and efficient ReID algorithms and establishing benchmarks covering various scenarios, such as traditional ReID [6, 52, 72, 74], clothes-changing ReID (CC-ReID) [13, 18, 21, 24, 46], clothes template based clothes-changing ReID (CTCC-ReID) [32, 64], visible-infrared ReID (VI-ReID) [36, 55, 56, 65] and text-to-image ReID (T2I-ReID) [3, 4, 30, 63, 76]. However, focusing solely on one specific scenario possesses inherent limitations. For instance, customers must deploy distinct models to retrieve persons according to the query information, which significantly increases the cost for model training and results in inconvenience for the application. To facilitate real-world deployment, there is a pressing need to devise one generic framework capable of re-identifying individuals across all scenarios mentioned above.

This paper proposes a new multi-purpose instruct-ReID task where existing **6** ReID settings can be formulated as its special cases. Specifically, the instruct-ReID uses query images and multi-model instructions as the model inputs, requiring the model to retrieve the same identity images from the gallery following the instructions. Using language or image instructions, models trained on the instruct-ReID task can be specialized to tackle diverse ReID tasks (Fig. 1a). For example, the clothes-changing ReID can be viewed as using the instruction "Ignore clothes" to retrieve. As another example, clothes template-based clothes-changing ReID can utilize a clothes template image as instruction. The proposed instruct-ReID task offers three significant advantages: easy deployment, improved performance, and easy extension to new ReID tasks. First, it enables cost-effective and convenient deployment in real-world applications. Unlike existing ReID approaches limited to specific tasks, instruct-ReID allows for utilizing a single model for various ReID scenarios, which is more practical in real applications. Second, as all existing ReID tasks can be consid-

*Corresponding authors.
†The work was done during an internship at SenseTime.



(a) One model for diverse ReID **(b) Improved Performance by OmniReID benchmark**

Figure 1. **(a)** We proposed a new instruct-ReID task that unites various ReID tasks. *Traditional ReID:* The instruction may be "Do not change clothes". *Clothes-changing ReID:* The instruction may be "Ignore clothes". *Clothes template based clothes-changing ReID:* The instruction is a cropped clothes image and the model should retrieve the same person wearing the provided clothing. *Language-instructed ReID:* The instruction is several sentences describing pedestrian attributes. The model is required to retrieve the person described by the instruction. *Visible-Infrared ReID:* The instruction can be "Cross modality". *Text-to-image ReID:* The model retrieves images according to the description sentence. **(b)** Our proposed method improves the performance of various person ReID tasks by a unified retrieval model.

ered special cases of instruct-ReID, we can unify the training sets of these tasks to exploit the benefits of more data and diverse annotations across various tasks – leading to enhanced performance. Third, instruct-ReID addresses a new language-instructed ReID task which requires the model to retrieve persons following language instructions. This setting is very practical in real applications for it enables customers to retrieve images they are particularly interested in through language descriptions. For example, customers can retrieve a woman wearing a black coat using one of her pictures and language instructions "Wear black coat".

To facilitate research in instruct-ReID, we introduce a new benchmark called OmniReID derived from 12 datasets * of 6 different ReID tasks. The OmniReID benchmark exhibits two appealing characteristics. First, it emphasizes ***diversity*** by incorporating images from various domains, including surveillance scenarios and synthetic games. The diversity ensures that the trained models are robust and can effectively handle ReID tasks in various real-world scenarios. Second, the OmniReID benchmark achieves ***comprehensiveness*** by offering evaluation datasets to assess various ReID tasks comprehensively, which facilitates evaluating the generalization ability of diverse ReID methods.

Based on OmniReID, we design a method with the proposed adaptive triplet loss for instruct-ReID. The typical triplet loss [44] only defines positive/negative pairs by identities, failing to align with instructions. Therefore, we propose a novel adaptive triplet loss to learn a metric space that preserves identity and instruction similarities. Specifically, we design an adaptive margin between two query-instruction pairs based on instruction similarities to pull features with similar instructions close and push features with different instructions apart. This loss incorporates the instruction information into the features representation and optimize the model to learn a metric space where similar instructions lead to closer features of query-instance pairs.

In summary, the contributions of this paper are three folds. (1) We propose a new instruct-ReID task, where existing traditional ReID, clothes-changing ReID, clothes template based clothes-changing ReID, visible-infrared ReID, text-to-image ReID andlanguage-instructed ReID can be viewed as special cases. (2) To facilitate research on instruct-ReID, we establish a large-scale and comprehensive OmniReID benchmark consisting 12 publicly available datasets. (3) We propose an adaptive triplet loss to supervise the feature distance of two query-instruction pairs to consider identity and instruction alignments. Our method consistently improves previous models on 10 datasets of 6 ReID tasks. For example, our method improves **+7.7%**, **+0.6%**, **+0.5%** mAP on CUHK03, MSMT17, Market1501 for traditional ReID, **+6.4%**, **+11.2%**, **+7.1%** mAP on PRCC, LTCC, VC-Clothes for clothes-changing ReID when using RGB images only, **+11.7%** mAP on COCAS+ real2 for clothes template based clothes changing ReID, **+4.3%** mAP on LLCM for visible-infrared ReID, **+2.6%** mAP on CUHK-PEDES for text-to-image ReID, **+24.9%** mAP on COCAS+ real2 for the new language-instructed ReID.

2. Related work

Person Re-identification. Person re-identification aims to retrieve the same images of the same identity with the given query from the gallery set. To support the ReID task on

*The discussion on ethical risks is provided in supplementary materials.

all-weather application, various tasks are conducted on the scenarios with changing environments, perspectives, and poses [4, 6, 62, 73, 76, 77]. Traditional ReID mainly focuses on dealing with indoor/outdoor problems when the target person wears the same clothes. To extend the application scenarios, clothes-changing ReID (CC-ReID) [64] and clothes template based clothes-changing ReID (CTCC-ReID) [32] are proposed. While CC-ReID forces the model to learn clothes-invariant features [13, 21], CCTC-ReID further extracts clothes template features [32] to retrieve the image of the person wearing template clothes. To capture person's information under low-light environments, visible-infrared person ReID (VI-ReID) methods [55, 56, 65, 65] retrieve the visible (infrared) images according to the corresponding infrared (visible) images. In the absence of the query image, GNA-RNN [30] introduced the text-to-image ReID (T2I-ReID) task, which aims at retrieving the person from the textual description. However, existing researches focus on a single scenario, making it difficult to address the demands of cross-scenario tasks. In this paper, we introduce a new Instruct-ReID task, which can be viewed as a superset of the existing ReID tasks by incorporating instruction information into identification.

Instruction Tuning. Instruction Tuning was first proposed to enable language models to execute specific tasks by following natural language instructions. Instruction-tuned models, *e.g.*, FLAN-T5 [7], InstuctGPT [41]/ChatGPT [40], UPT [17] can effectively prompt the ability on zero- and few-shot transfer tasks. A few works borrowed the idea from language to vision. Flamingo [2], BLIP-2 [27], and KOSMOS-1 [20] learning with image-text pairs also show promising generalization on visual understanding tasks. While these methods aim to generate convincing language responses following the image or language instructions, we focus on retrieving the correct person following the given instructions by tuning a vision transformer.

Multi-model Retrieval. Multi-model retrieval is widely used to align information from multiple modalities and improve the performance of applications. In multi-model retrieval, unimodal encoders always encode different modalities for retrieval tasks. For instance, CLIP [43], VideoCLIP [54], COOT [12] and MMV [1] utilize contrastive learning for pre-training. Other techniques like HERO [28], Clipbert [25], Vlm [53], TAM [39] and MVLT [22] focus on merging different modalities for retrieval tasks to learn a generic representation. Although there have been numerous studies on multi-modal retrieval, most are concentrated on language-vision pretraining or video retrieval, leaving the potential of multi-modal retrieval for person ReID largely unexplored. This paper aims to investigate this underexplored area to retrieve anyone with information extracted from multiple modalities.

Table 1. Comparison of training subsets of different ReID datasets.

dataset	image	ID	domain
MSMT17 [52]	30,248	1,041	indoor/outdoor
Market1501 [71]	12,936	751	outdoor
PRCC [57]	17,896	150	indoor
COCAS+ Real1 [32]	34,469	2,800	indoor/outdoor
LLCM [67]	30,921	713	indoor/outdoor
LaST [47]	71,248	5,000	indoor/outdoor
MALS [58]	1,510,330	1,510,330	synthesis
LUPerson-T [45]	957,606	-	indoor/outdoor
OmniReID	4,973,044	328,604	indoor/outdoor/synthetic

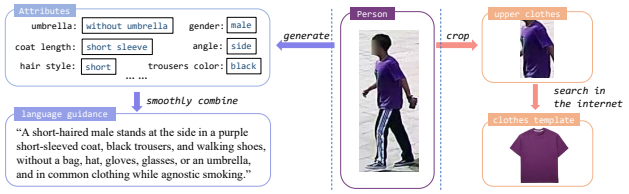
3. OmniReID Benchmark

To facilitate research on Instruct-ReID, we propose the OmniReID benchmark including a large-scale pretraining dataset based on 12 publicly available datasets with visual and language annotations. The comparison with the existing ReID benchmark is illustrated in Tab. 1.

Protocols. To enable all-purpose person ReID, we collect massive public datasets from various domains and use their training subset as our training subset, including Market1501 [71], MSMT17 [52], CUHK03 [34] for traditional ReID, PRCC [57], VC-Clothes [50], LTCC [42] for clothes-changing ReID, LLCM [67] for visible-infrared ReID, CUHK-PEDES [30], SYNTH-PEDES [79] for text-to-image ReID, COCAS+ Real1 [67] for clothes template based clothes-changing ReID and language-instructed ReID, forming 4,973,044 images and 328,604 identities. To fairly compare our method with state-of-the-art methods, the trained models are evaluated on LTCC, PRCC, VC-Clthoes, COCAS+ Real2, LLCM, CUHK-PEDES, Market1501, MSMT17, and CUHK03 test subsets without fine-tuning. We search the target images with query images and the instructions. Since a query image has multiple target images in the gallery set and CMC (Cumulative Matching Characteristic) curve only reflects the retrieval precision of most similar target images, we also adopt mAP (mean Average Precision) to reflect the overall ranking performance w.r.t. all target images. We present all dataset statistics of OmniReID in the supplementary materials.

Language Annotation Generation. To generate language instruction for language-instructed ReID, we annotate COCAS+ Real1 and COCAS+ Real2 with language description labels. Similar to Text-to-Image ReID datasets, language annotations in OmniReID are several sentences that describe the visual appearance of pedestrians. We divide our annotation process into *pedestrian attribute generation* and *attribution-to-language transformation*.

Pedestrian Attribute Generation. To obtain a comprehensive and varied description of an individual, we employ an extensive collection of attribute words describing a wide range of human visual characteristics. The collection contains 20 attributes and 92 specific representation words, including full-body clothing, hair color, hairstyle, gender,



(a) language guidance generation **(b) clothes template collection**

Figure 2. **(a)** We generate attributes for a person and then transform attributes into sentences by a large language model. **(b)** We crop upper clothes and search them online for clothes templates.

posture, and accessories such as umbrellas or satchels. Professional annotators manually label all the pedestrian attributes. We provide a practical illustration with the attribute collection in Fig. 2(a). By utilizing instructions on the well-defined attribute combination, models can further enhance their ability to identify the target person.

Attribute-to-Language Transformation. Compared with discrete attribute words, language is more natural for consumers. To this end, we transform these attributes into multiple sentences using the Alpaca-LoRA [66] large language model. Specifically, we ask the Alpaca-LoRA with the following sentences: "Generate sentences to describe a person. The above sentences should contain all the attribute information I gave you in the following." The generated annotations are carefully checked and corrected manually to ensure the correctness of the language instructions. All detailed pedestrian attributes and more language annotations are presented in the supplementary materials.

Visual Annotation Generation. Visual annotations are images that describe the characteristics of pedestrians. In this paper, we select clothes as the visual annotations because they are viewed as the most significant visual characteristics of pedestrians. To get high-quality visual annotations, we first crop the upper clothes from the source images (Clothes Copping) and search on the internet to get the most corresponding clothes-template images (Clothes-template Crawling) as visual annotations. Since each person wears the same clothes in traditional ReID datasets, we annotate the clothes-changing LTCC dataset where each person wears multiple clothes to ease the burden of annotations. Fig. 2 (b) shows the detailed process.

Clothes Cropping. We use a human parsing model SCHP [29] to generate the segmentation mask of the upper clothes and then crop the corresponding rectangle upper clothes patches from the original images. These bounding boxes of upper clothes are then manually validated.

Clothes-template Crawling. Given all cropped upper clothes from images in OmniReID datasets, we crawl the templates of these clothes from shopping websites[†]. The top 40 matching clothes templates are downloaded when we crawl each cropped upper clothes. The one with the highest image quality is manually selected.

4. IRM: Instruct ReID Model

As shown in Fig. 3 (a), the proposed Instruct ReID model consists of four parts: instruction generation (Sec. 4.1) for various ReID tasks, an editing transformer $\mathcal{E}_e$ (Sec. 4.2) and an instruction encoder $\mathcal{E}_i$ which is a visual language model, and a cross-model attention module $\mathcal{E}_f$. Given an instruction $\mathbf{T}$ associated with a query image $\mathbf{I}$, we obtain instruction features $\mathbf{F_T}$ using the instruction encoder $\mathcal{E}i$. These extracted instruction features $\mathbf{F}_T$, along with query image tokens, are fed into the designed editing transformer $\mathcal{E}_e$ to obtain features $\mathbf{F}$ edited by instructions. Furthermore, we introduce an attention module $\mathcal{E}_f$ to efficiently combine features of the query image and instruction through the attention mechanism in the stacked transformer layers. Finally, the overall loss function is introduced (Sec. 4.4) for all ReID tasks, which includes our newly proposed adaptive triplet loss $\mathcal{L}_{atri}$ (Sec. 4.3) instead of traditional triplet loss.

4.1. Instruction Generation

In our proposed instruct-ReID task, the model must retrieve the images that describe the same person following the instructions. By designing different instructions, our instruct-ReID can be specialized to existing ReID tasks, *i.e.,* traditional ReID, clothes-changing ReID, clothes template based clothes-changing ReID, visual-infrared ReID, text-to-image ReID, and language-instructed ReID. We show the current instructions and leave the exploration of better instructions for instruct-ReID to future research.

Traditional ReID: Following Instruct-BLIP [10], we generate 20 instructions[‡] from GPT-4 and randomly select one, *e.g.,* "Do not change clothes.", when training. The model is expected to retrieve images of the same person without altering image attributes, such as clothing.

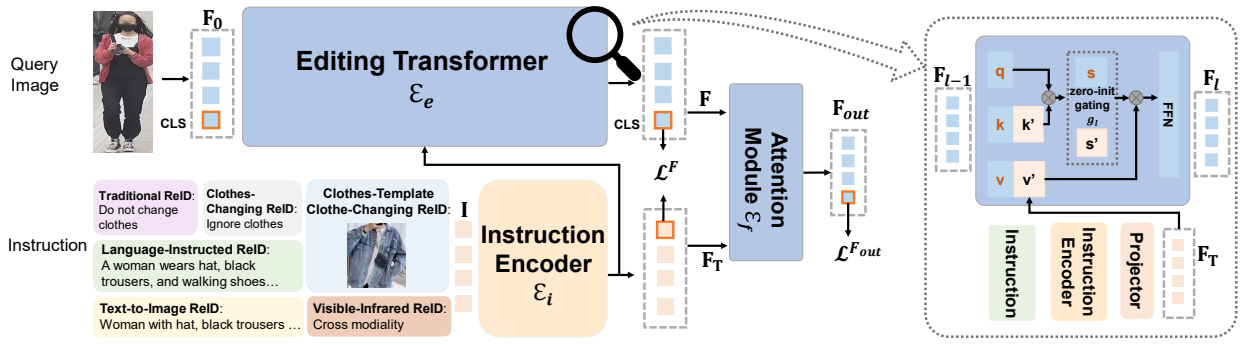
> ### **Query image**:{Query image}
> ### **Instruction**: "Do not change clothes."[3]
> ### **Target image**:{Output}

Clothes-changing ReID: Simiar to Traditional ReID, the instruction is the sentence chosen from a collection of 20 GPT-4 generated sentences[3], *e.g.,* "With clothes changed". The model should retrieve images of the same person even when wearing different outfits.

> ### **Query image**:{Query image}
> ### **Instruction**: "With clothes changed."[3]
> ### **Target image**:{Output}

[†]https://world.taobao.com/, https://www.17qcc.com/

[‡]See all available instruction sentences in supplementary materials.



(a) Instruct-ReID Model **(b) l-th editing transformer**

Figure 3. The overall architecture of the proposed method. The instruction is fed into the instruction encoder to extract instruction features (a). The features are then propagated into the editing transformer (b) to capture instruction-edited features. We exploit adaptive triplet loss and identification loss to train the network. For the testing stage, we use the CLS token for retrieval.

Clothes template based clothes-changing ReID: The instruction is a clothes template for a query image while a cropped clothes image for a target image. The model should retrieve images of the same person wearing the provided clothes. We provide more examples for training and test in the supplementary materials.

> ### **Query image**:{Query image}
> ### **Instruction**:{Any clothes template image}
> ### **Target image**:{Output}

Visual-Infrared ReID: The instruction is the sentence chosen from a collection of 20 GPT-4 generated sentences[3], *e.g.,* "Retrieve cross modality image". The model should retrieve visible (infrared) images of the same person according to the corresponding infrared (visible) images.

> ### **Query image**:{Query image}
> ### **Instruction**: "Retrieve cross modality image."[3]
> ### **Target image**:{Output}

Text-to-Image ReID: The instruction is the describing sentences, and both images and text are fed into IRM during the training process. While in the inference stage, the image features and instruction features are extracted separately. [§]

> ### **Image**:{Image}[4]
> ### **Instruction**:{Sentences describing pedestrians}[4]
> ### **Target image**:{Output}

Language-instructed ReID: The instruction is several sentences describing pedestrian attributes. We randomly select the description languages from the person images in gallery and provide to query images as instruction. The model is required to retrieve images of the same person described in the provided sentences. We provide more examples for training and test in the supplementary materials.

[§]Image and instruction features are extracted separately in test stage.

> ### **Query image**:{Query image}
> ### **Instruction**:{Sentences describing pedestrians}
> ### **Target image**:{Output}

4.2. Editing Transformer

The editing transformer consists of L zero-init transformer layers $\mathcal{E}_e = \{\mathcal{F}_1, \mathcal{F}_2, ..., \mathcal{F}_L\}$, which can leverage the instruction to edit the feature of query images. Given the l-th zero-init transformer, the output feature $\mathbf{F}_l$ can be formulated as

$$\mathbf{F}_l = \mathcal{F}_l(\mathbf{F}_{l-1}, \mathbf{F_T}), \tag{1}$$

where $\mathbf{F_T} = \mathcal{E}_i(\mathbf{I})$ is the instruction feature extracted by $\mathcal{E}_i$ and $\mathbf{F}_{l-1}$ is the output feature of $(l-1)$-th zero-init transformer layer. The initial input $\mathcal{F}_0$ of the first layer is defined as $\mathbf{F}_0 = [\mathbf{f}_0^{\text{CLS}}, \mathbf{f}_0^1, \mathbf{f}_0^2, ..., \mathbf{f}_0^N]$, where $\mathbf{f}_0^{\text{CLS}}$ is the [CLS] token, $(\mathbf{f}_0^1, \mathbf{f}_0^2, ..., \mathbf{f}_0^N)$ are the patch tokens of the query image and N is the number of patches of the query image.

We show the detailed structure of each layer in the editing transformer in Fig. 3 (b). Given the features $\mathbf{F}_{l-1}$ and instruction features $\mathbf{F_T}$, the attention map $\mathbf{M}_l$ is defined as

$$\mathbf{M}_l = \left[\text{Softmax}(\mathbf{S}_l), g_l \times \text{Softmax}(\mathbf{S'}_l)\right], \tag{2}$$

where g_l is the gating parameters initialized by zero. Here, $\mathbf{S}_l$ is the attention map between queries and keys of input features and $\mathbf{S}'_l$ is the attention map between queries of input features and keys of instruction features. Mathematically,

$$\mathbf{S}_l = \mathbf{Q}_l\mathbf{K}_l^\top/\sqrt{C}, \mathbf{S}'_l = \mathbf{Q}_l\mathbf{K'}_l^\top/\sqrt{C}, \tag{3}$$

where a linear projection derives queries and keys, *i.e.,* $\mathbf{Q}_l = \text{Linear}_q(\mathbf{F}_{l-1})$, $\mathbf{K}_l = \text{Linear}_k(\mathbf{F}_{l-1})$ and $\mathbf{K}'_l = \text{Linear}_{k'}(\mathbf{F_T})$, respectively. C is the feature dimension of query features. Finally, we calculate the output of the l-th layer by

$$\mathbf{F}_l = \text{Linear}_o(\mathbf{M}_l\left[\mathbf{V}_l, \mathbf{V}'_l\right]), \tag{4}$$

where Linear_o is the feed-forward network after the attention layer in each transformer block, $\mathbf{V}_l$ and $\mathbf{V}'_l$ are

the values calculated by $\mathbf{V}_l = \text{Linear}_v(\mathbf{F}_{l-1})$ and $\mathbf{V}'_l = \text{Linear}_{v'}(\mathbf{F_T})$. We use the [CLS] token in the output feature of L-th transformer layer for computing losses and retrieval, *i.e.,* $\mathbf{F} = \mathbf{f}_L^{\text{CLS}}$, where $\mathbf{F}_L = (\mathbf{f}_L^{\text{CLS}}, \mathbf{f}_L^1, \mathbf{f}_L^2, ..., \mathbf{f}_L^N)$ and N is patch number of query images.

4.3. Adaptive Triplet Loss

Unlike typical triplet loss that defines positive and negative samples solely based on identities, instruct-ReID requires distinguishing images with different instructions for the same identity. Intuitively, an adaptive margin should be set to push or pull samples based on the instruction difference. Let $(\mathbf{F}_i^a, \mathbf{F}_i^{r_1}, \mathbf{F}_i^{r_2})$ be the i-th triplet in the current mini-batch, where $\mathbf{F}_i^a$ is an anchor sample, $\mathbf{F}_i^{r_1}$ and $\mathbf{F}_i^{r_2}$ are reference samples. We propose an adaptive triplet loss as

$$\mathcal{L}_{atri} = \frac{1}{N_{tri\dagger}} \sum_{i=1}^{N_{tri\dagger}} \{\mathbf{Sign}(\beta_1 - \beta_2)[d(\mathbf{F}_i^a, \mathbf{F}_i^{r_1}) + (\beta_1 - \beta_2)\, m - d(\mathbf{F}_i^a, \mathbf{F}_i^{r_2})]\}_+ \quad (5)$$

where $N_{tri\dagger}$ and m denote the number of triplets and a hyper-parameter for the maximal margin, respectively. d is a Euclidean distance function, *i.e.*, $d(\mathbf{F}_i^a, \mathbf{F}_i^r) = ||\mathbf{F}_i^a - \mathbf{F}_i^r||_2^2$. β_1 and β_2 are relatednesses between the anchor image and the corresponding reference image that consider the identity consistency and instruction similarity for the adaptive margin. Mathematically, they are defined as

$$\beta_j = \mathbb{I}\left(y_a = y_{r_j}\right) \mathbf{Cos}\left\langle \mathbf{F}_{\mathbf{T}}^a, \mathbf{F}_{\mathbf{T}}^{r_j} \right\rangle, \quad (6)$$

where y_a and y_{r_j} are the identity labels of the anchor image and the reference image, $\mathbb{I}(\cdot)$ is the indicator function, and $j = \{1, 2\}$ denotes the index of reference samples.

The concept of adaptive triplet loss is described by Fig. 4(a). We discuss the adaptive loss in two cases. First, as shown in Fig. 4(b), the margin is set to zero if the triplet has the same identity and the instructions of the two reference samples are equally similar to the instruction of the anchor sample. This makes the distances between the anchor point and the two reference points the same. Second, as shown in Fig. 4(c), when there is a significant difference in instruction similarities, the margin distance between the anchor and two references becomes closer to the maximum value m, forcing the model to learn discriminative features. Adaptive triplet loss makes features from the same person become distinctive based on the similarity of instructions, which helps to retrieve images that align the requirements of given instructions in the CTCC-ReID and LI-ReID tasks.

4.4. Overall Loss Function

We impose an identification loss $\mathcal{L}_{id}$, which is the classification loss on identities, and an adaptive triplet loss $\mathcal{L}_{atri}$ on $\mathbf{F}$ and the final fusion features $\mathbf{F_{out}}$ to supervise the model

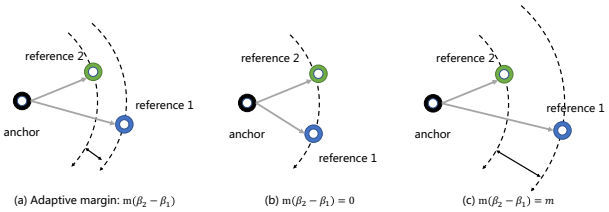


Figure 4. Illustration of adaptive triplet loss. Unlike the traditional triplet loss where the margin is fixed, the margin in our adaptive triplet loss is defined by the instruction similarity for the two query-instruction pairs that describe the same person. The features associated with similar instructions are pulled to be closer.

for training Trad-ReID, CC-ReID, VI-ReID, CTCC-ReID, and LI-ReID tasks. The overall loss is combined as

$$\mathcal{L} = \mathcal{L}_{atri}(F) + \mathcal{L}_{id}(F) + \mathcal{L}_{atri}(F_{out}) + \mathcal{L}_{id}(F_{out}) \quad (7)$$

where $\mathbf{F}$ and $\mathbf{F}_{out}$ indicate the source of features used in calculating the loss.

For the T2I-ReID task, we adopt a contrastive loss $\mathcal{L}_{cl}$ to align the image features $\mathbf{F}$ and text features $\mathbf{F_T}$ to enable text-based person retrieval. We also employ a binary classification loss $\mathcal{L}_{match}$ to learn whether an inputted image-text pair is positive or negative, defined as:

$$\mathcal{L} = \mathcal{L}_{cl}(F) + \mathcal{L}_{match}(F_{out}) \quad (8)$$

where, $\mathcal{L}_{match}(F_{out}) = \mathcal{C}_e(\hat{y}, F_{out}) = \mathcal{C}_e[\hat{y}, \mathcal{E}_f(\mathbf{F}, \mathbf{F_T})]$, respectively. $\mathcal{C}_e$ is a binary cross-entropy loss, $\hat{y}$ is a 2-dimension one-hot vector representing the ground-truth label *i.e.,* $[0, 1]^\intercal$ for positive pair and $[1, 0]^\intercal$ for negative pair, formed by matching text features with corresponding image features before inputting into the attention module $\mathcal{E}_f$.

5. Experiment

5.1. Experimental Setups

Training Settings. To enable all-purpose person ReID, we perform two training scenarios based on the built benchmark **OmniReID**: 1) Single-task Learning (STL): Every task is trained and tested individually using the corresponding dataset. 2) Multi-task Learning (MTL): To acquire one unified model for all tasks, the model is optimized by joint training of the four ReID tasks with all the training datasets. The trained network is then evaluated for different tasks on various datasets. We provide the details of datasets used in STL and MTL in the supplementary materials.

Implementation Details. For the editing transformer, we use the plain ViT-Base with the ReID-specific pretraining [78]. The instruction encoder is ALBEF [26]. All images are resized into 256×128 for training and test. We use the AdamW optimizer with a base learning rate of 1e-5 and a weight decay of 5e-4 . We linearly warmup the learning rate from 1e-7 to 1e-5 for the first 1000 iterations. Random

Table 2. The performance of Clothes-Changing ReID of our method and the state-of-the-art methods. Mean average precision (mAP) and Top1 are used to quantify the accuracy. † denotes that the model is trained with multiple datasets. * denotes that the model is pre-trained on 4 million images.

Methods	LTCC		PRCC		VC-Clothes	
	mAP	Top1	mAP	Top1	mAP	Top1
HACNN[33]	26.7	60.2	-	21.8	-	-
RGA-SC[69]	27.5	65.0	-	42.3	67.4	71.1
PCB[48]	30.6	65.1	38.7	41.8	62.2	62.0
IANet[19]	31.0	63.7	45.9	46.3	-	-
CAL[13]	40.8	74.2	-	-	-	-
TransReID[16]	-	-	-	44.2	71.8	72.0
IRM (STL)	46.7	66.7	46.0	48.1	**80.1**	**90.1**
IRM (MTL)†	**52.0**	**75.8**	**52.3**	**54.2**	78.9	89.7

Table 3. Comparison with the state-of-the-art methods on visible-infrared ReID and text-to-image ReID. The VI-ReID setting is *VIS-to-IR* and *IR-to-VIS* in LLCM. † denotes that the model is trained with multiple datasets.* denotes that the model is trained under the same image shape as IRM, *i.e.,* 256×128.

Methods	T2I-ReID		VI-ReID: LLCM			
	CUHK-PEDES		VIS-to-IR		IR-to-VIS	
	mAP	Top1	mAP	Top1	mAP	Top1
ALBEF [26]	56.7	60.3	-	-	-	-
CAJ [61]	-	-	59.8	56.5	56.6	48.8
MMN [68]	-	-	62.7	59.9	58.9	52.5
DEEN [67]	-	-	65.8	62.5	62.9	54.9
SAF [31]	58.6	64.1	-	-	-	-
PSLD [15]	60.1	64.1	-	-	-	-
RaSa* [3]	63.9	69.6	-	-	-	-
IRM (STL)	65.3	72.8	66.6	66.2	64.5	64.9
IRM (MTL)†	**66.5**	**74.2**	**67.5**	**66.7**	**67.2**	**65.7**

cropping, flipping, and erasing are used for data augmentation during training. For each training batch, we randomly select 32 identities with 4 image samples for each identity.

5.2. Experimental Results

Clothes-Changing ReID (CC-ReID). As shown in Tab 2, IRM outperforms all state-of-the-art methods on LTCC, PRCC and VC-Clothes, showing that the model can effectively extract clothes-invariant features following the instructions, *e.g.,* "Ignore clothes". Specifically, on STL, IRM improves CAL [13], TransReID [16] by **+5.9%** mAP and **+8.3%** mAP on LTCC and VC-Clothes, respectively. On MTL, IRM further improves the performance on LTCC to **52.0%** mAP and reaches a new state-of-the-art on PRCC with **52.3%** mAP. Other methods like ACID [60], CCFA [14], AIM [59] and DCR-ReID [9] provide results of 384×192 image size, we present more results based on this resolution of IRM in supplementary. While multi-task learning leads to slightly lower performance than single-task learning on VC-Clothes, our method still achieves a higher Top-1 than TransReID. We conjecture that this drop is due to the domain gap between VC-Clothes (Synthetic) and other datasets (Real) and leave it for future work.

Clothes Template Based Clothes-Changing ReID (CTCC-ReID). Our method achieves desirable performance on the CTCC-ReID task in Tabel 4, which shows that a fixed instruction encoder is enough for this tough task. Concretely, when only trained on COCAS+ Real1, IRM outperforms BC-Net [64] and DualBCT-Net [32], both of which learn an independent clothes branch, by **+9.6%** mAP and **+2.2%** mAP, respectively. By integrating the knowledge on other instruct ReID tasks during multi-task learning, we are able to further improve the performance of IRM, achieving an mAP of **41.7%** and pushing the performance limits on CTCC-ReID.

Visible-Infrared ReID (VI-ReID). We validate the performance of IRM on Visible-Infrared ReID datasets LLCM, which is a new and challenging low-light cross-modality dataset and has a more significant number of identities and images. From Tab. 3, we can see that the results on the two test modes show that the proposed IRM achieves competitive performance against all other state-of-the-art methods. Specifically, for the VIS-to-IR mode on LLCM, IRM achieves **67.5%** mAP and exceeds previous state-of-the-art methods like DEEN [67] by **+1.7%**. For the IR-to-VIS mode on LLCM, IRM achieves **65.7%** Rank-1 accuracy and **67.2%** mAP, which is a new state-of-the-art result. The results validate the effectiveness of our method.

Text-to-Image ReID (T2I-ReID). As shown in Tab. 3, IRM shows competitive performance with a mAP of **66.5%** on the CUHK-PEDES [30], which is **+2.6%**, **+6.4%**, **+7.9%** higher than previous methods RaSa [3] (63.9%), PSLD [15] (60.1%), SAF [31] (58.6%). Because a few images in CUHK-PEDES training set are from the test sets of Market1501 and CUHK03, we filtered out duplicate images from the test sets during the multi-task learning (MTL) process. The testing was conducted on uniformly resized images with a resolution of 256×128.

Language-Instructed ReID (LI-ReID). As a new setting in ReID, no previous works have been done to retrieve a person using several sentences as the instruction, therefore, we compare IRM with a straightforward baseline. In the baseline method, only a ViT-Base is trained on COCAS+ Real1 images without utilizing the information from language instruction, leading to poor person re-identification ability. As shown in Tab 4, IRM improves the baseline by **+15.8%** mAP, because IRM can integrate instruction information into identity features. With MTL, IRM achieves extra **+9.1%** performance gain by using more images and general information in diverse ReID tasks.

Traditional ReID (Trad-ReID). IRM also shows its effectiveness on Trad-ReID tasks in Tab 4. Specifically,

Table 4. Performance comparison with the state-of-the-art methods on Clothes -Template Clothes-Changing ReID, Language-Instructed ReID, and Traditional ReID. † denotes that the model is trained with multiple datasets. * denotes that the model is pretrained on 4 million pedestrian images. The size of the input images used in the table is 256x128.

Methods	CTCC-ReID		LI-ReID		Trad-ReID					
	COCAS+ Real2		COCAS+ Real2		Market1501		MSMT17		CUHK03	
	mAP	Top1	mAP	Top1	mAP	Top1	mAP	Top1	mAP	Top1
Baseline	-	-	14.9	31.6	-	-	-	-	-	-
TransReID [16]	5.5	17.5	-	-	86.8	94.4	61.0	81.8	-	-
BC-Net [64]	22.6	36.9	-	-	-	-	-	-	-	-
DualBCT-Net [32]	30.0	48.9	-	-	-	-	-	-	-	-
SAN [23]	-	-	-	-	88.0	96.1	-	-	76.4	80.1
HumanBench† [49]	-	-	-	-	89.5	-	69.1	-	77.7	-
PASS* [78]	-	-	-	-	93.0	96.8	71.8	88.2	-	-
IRM (STL)	32.2	54.8	30.7	60.8	92.3	96.2	71.9	86.2	83.3	86.5
IRM (MTL)†	**41.7**	**64.9**	**39.8**	**71.6**	**93.5**	**96.5**	**72.4**	**86.9**	**85.4**	**86.5**

Table 5. Ablation study. The performance comparison of our editing transformer and using ViT base transformer (w/o editing), and the comparisons with triplet loss (w/o $\mathcal{L}_{atri}$) and the proposed adaptive triplet loss in terms of mAP. † denotes that the test mode is VIS-to-IR and ‡ denotes IR-to-VIS mode on LLCM.

Methods	CTCC-ReID	LI-ReID	T2I-ReID	VI-ReID		CC-ReID			Trad-ReID			Avg.
	Real2	Real2	CUHK.	LLCM†	LLCM‡	LTCC	PRCC	VC-Clo.	Market1501	MSMT17	CUHK03	
IRM (STL)	32.2	30.7	65.3	66.6	64.5	46.7	46.0	80.1	92.3	71.9	83.3	61.8
w/o editing	32.6	30.5	65.7	66.2	65.1	46.2	45.7	77.8	92.5	71.4	83.2	61.5
w/o $\mathcal{L}_{atri}$	31.5	30.2	-	66.6	64.5	46.7	46.0	80.1	92.3	71.9	83.3	61.3
IRM (MTL)	41.7	39.8	66.5	67.5	67.2	52.0	52.3	78.9	93.5	72.4	85.4	65.1
w/o editing	41.0	38.8	66.1	67.3	65.2	51.2	51.0	78.6	93.0	71.8	85.1	64.5
w/o $\mathcal{L}_{atri}$	40.7	39.2	65.2	68.4	66.3	52.0	52.4	77.9	92.9	72.0	85.5	64.8

when trained on a single dataset, compared with PASS [78], IRM achieves comparable performance on Market1501, MSMT17 and **+6.9%** performance gain on CUHK03 compared with SAN [23]. With multi-task training, IRM can outperform the recent multi-task pretraining HumanBench [49] and self-supervised pretraining [78]. We do not compare with SOLDIER [5] because it only reports ReID results with the image size of 384x192 instead of 256x128 in our method. More importantly, SOLDIER primarily focuses on pretraining and is evaluated on traditional ReID tasks only, while the claimed contribution of IRM is to tackle multiple ReID tasks with one model.

5.3. Ablation Study

Editing Transformer. To verify the effectiveness of the instruction integrating design in the editing transformer, we compare it with the traditional ViT base model, where the image features are extracted without fusing information from instruction features. Results in Tab 5 show that adopting the editing transformer leads to **-0.6%** mAP performance drop in the MTL scenario. Consistent results can be observed in STL, indicating the effectiveness of the instruction integrating design in the editing transformer.

Adaptive Triplet Loss. Tab 5 shows that adaptive triplet loss in MTL outperforms the traditional triplet loss by **+0.3%** mAP on average, indicating that the proposed loss boosts the model to learn more discriminative features following different instructions. On STL, for CC-ReID and Trad-ReID, the instructions are fixed sentences leading to the same performance of adaptive/traditional triplet loss. However, in the case of CTCC-ReID and LI-ReID, where instructions vary among samples, using adaptive triplet loss brings about **+0.7%**, **+0.5%** mAP gain, which shows the effectiveness of adaptive triplet loss in learning both identity and instruction similarity.

More Results. We provide additional results based on different pre-trained models and visualizations of the retrieval results in the supplementary material.

6. Conclusion

This proposes one unified instruct-ReID task to jointly tackle existing traditional ReID, clothes-changing ReID, clothes template based clothes-changing ReID, language-instruct ReID, visual-infrared ReID, and text-to-image ReID tasks, which holds great potential in social surveillance. To tackle the instruct-ReID task, we build a large-scale and comprehensive OmniReID benchmark and a generic framework with an adaptive triplet loss. We hope our OmniReID can facilitate future works such as unified network structure design and multi-task learning methods on a broad variety of retrieval tasks.

Acknowledgement.This paper was supported by the National Natural Science Foundation of China (No.62127803), Key R&D Project of Zhejiang Province (No.2022C01056).

References

[1] Jean-Baptiste Alayrac, Adria Recasens, Rosalia Schneider, Relja Arandjelović, Jason Ramapuram, Jeffrey De Fauw, Lucas Smaira, Sander Dieleman, and Andrew Zisserman. Self-supervised multimodal versatile networks. *NeurIPS*, 2020. 3

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *NeurIPS*, 2022. 3

[3] Yang Bai, Min Cao, Daming Gao, Ziqiang Cao, Chen Chen, Zhenfeng Fan, Liqiang Nie, and Min Zhang. Rasa: Relation and sensitivity aware representation learning for text-based person search. *arXiv preprint arXiv:2305.13653*, 2023. 1, 7

[4] Dapeng Chen, Hongsheng Li, Xihui Liu, Yantao Shen, Jing Shao, Zejian Yuan, and Xiaogang Wang. Improving deep visual representation for person re-identification by global and local image-language association. In *ECCV*, 2018. 1, 3

[5] Weihua Chen, Xianzhe Xu, Jian Jia, Hao Luo, Yaohua Wang, Fan Wang, Rong Jin, and Xiuyu Sun. Beyond appearance: A semantic controllable self-supervised learning framework for human-centric visual tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15050–15061, 2023. 8

[6] Ying-Cong Chen, Xiatian Zhu, Wei-Shi Zheng, and Jian-Huang Lai. Person re-identification by camera correlation aware feature augmentation. *TPAMI*, 2017. 1, 3

[7] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022. 3

[8] Yuanzheng Ci, Yizhou Wang, Meilin Chen, Shixiang Tang, Lei Bai, Feng Zhu, Rui Zhao, Fengwei Yu, Donglian Qi, and Wanli Ouyang. Unihcp: A unified model for human-centric perceptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17840–17852, 2023. 1

[9] Zhenyu Cui, Jiahuan Zhou, Yuxin Peng, Shiliang Zhang, and Yaowei Wang. Dcr-reid: Deep component reconstruction for cloth-changing person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023. 7

[10] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023. 4

[11] Niloofar Gheissari, Thomas B Sebastian, and Richard Hartley. Person reidentification using spatiotemporal appearance. In *CVPR*, 2006. 1

[12] Simon Ging, Mohammadreza Zolfaghari, Hamed Pirsiavash, and Thomas Brox. Coot: Cooperative hierarchical transformer for video-text representation learning. *NeurIPS*, 2020. 3

[13] Xinqian Gu, Hong Chang, Bingpeng Ma, Shutao Bai, Shiguang Shan, and Xilin Chen. Clothes-changing person re-identification with rgb modality only. In *CVPR*, 2022. 1, 3, 7

[14] Ke Han, Shaogang Gong, Yan Huang, Liang Wang, and Tieniu Tan. Clothing-change feature augmentation for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22066–22075, 2023. 7

[15] Xiao Han, Sen He, Li Zhang, and Tao Xiang. Text-based person search with limited data. *arXiv preprint arXiv:2110.10807*, 2021. 7

[16] Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. Transreid: Transformer-based object re-identification. In *CVPR*, 2021. 7, 8

[17] Weizhen He, Weijie Chen, Binbin Chen, Shicai Yang, Di Xie, Luojun Lin, Donglian Qi, and Yueting Zhuang. Unsupervised prompt tuning for text-driven object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2651–2661, 2023. 3

[18] Peixian Hong, Tao Wu, Ancong Wu, Xintong Han, and Wei-Shi Zheng. Fine-grained shape-appearance mutual learning for cloth-changing person re-identification. In *CVPR*, 2021. 1

[19] Ruibing Hou, Bingpeng Ma, Hong Chang, Xinqian Gu, Shiguang Shan, and Xilin Chen. Interaction-and-aggregation network for person re-identification. In *CVPR*, 2019. 7

[20] Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Qiang Liu, et al. Language is not all you need: Aligning perception with language models. *arXiv preprint arXiv:2302.14045*, 2023. 3

[21] Yan Huang, Qiang Wu, JingSong Xu, Yi Zhong, and ZhaoXiang Zhang. Clothing status awareness for long-term person re-identification. In *CVPR*, 2021. 1, 3

[22] Ge-Peng Ji, Mingchen Zhuge, Dehong Gao, Deng-Ping Fan, Christos Sakaridis, and Luc Van Gool. Masked vision-language transformer in fashion. *Machine Intelligence Research*, 20(3):421–434, 2023. 3

[23] Xin Jin, Cuiling Lan, Wenjun Zeng, Guoqiang Wei, and Zhibo Chen. Semantics-aligned representation learning for person re-identification. In *AAAI*, 2020. 8

[24] Xin Jin, Tianyu He, Kecheng Zheng, Zhiheng Yin, Xu Shen, Zhen Huang, Ruoyu Feng, Jianqiang Huang, Zhibo Chen, and Xian-Sheng Hua. Cloth-changing person re-identification from a single image with gait prediction and regularization. In *CVPR*, 2022. 1

[25] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. Less is more: Clipbert for video-and-language learning via sparse sampling. In *CVPR*, 2021. 3

[26] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021. 6, 7

[27] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 3

[28] Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. Hero: Hierarchical encoder for video+ language omni-representation pre-training. *arXiv preprint arXiv:2005.00200*, 2020. 3

[29] Peike Li, Yunqiu Xu, Yunchao Wei, and Yi Yang. Self-correction for human parsing. *TPAMI*, 2020. 4

[30] Shuang Li, Tong Xiao, Hongsheng Li, Bolei Zhou, Dayu Yue, and Xiaogang Wang. Person search with natural language description. In *CVPR*, 2017. 1, 3, 7

[31] Shiping Li, Min Cao, and Min Zhang. Learning semantic-aligned feature representation for text-based person search. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2724–2728. IEEE, 2022. 7

[32] Shihua Li, Haobin Chen, Shijie Yu, Zhiqun He, Feng Zhu, Rui Zhao, Jie Chen, and Yu Qiao. Cocas+: Large-scale clothes-changing person re-identification with clothes templates. *TCSVT*, 2022. 1, 3, 7, 8

[33] Wei Li and Shaogang Gong. Harmonious attention network for person re-identification. In *CVPR*, 2018. 7

[34] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, 2014. 3

[35] Zhiqiu Lin, Samuel Yu, Zhiyi Kuang, Deepak Pathak, and Deva Ramana. Multimodality helps unimodality: Cross-modal few-shot learning with multimodal models. *arXiv preprint arXiv:2301.06267*, 2023. 1

[36] Jialun Liu, Yifan Sun, Feng Zhu, Hongbin Pei, Yi Yang, and Wenhui Li. Learning memory-augmented unidirectional metrics for cross-modality person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19366–19375, 2022. 1

[37] Haoyu Lu, Yuqi Huo, Mingyu Ding, Nanyi Fei, and Zhiwu Lu. Cross-modal contrastive learning for generalizable and efficient image-text retrieval. *Machine Intelligence Research*, 20(4):569–582, 2023. 1

[38] Bence Nanay. Multimodal mental imagery. *Cortex*, 2018. 1

[39] Xuesong Nie, Xi Chen, Haoyuan Jin, Zhihang Zhu, Yunfeng Yan, and Donglian Qi. Triplet attention transformer for spatiotemporal predictive learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 7036–7045, 2024. 3

[40] OpenAI. Chatgpt. Available at https://openai.com/blog/chatgpt/, 2023. 3

[41] 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*, 2022. 3

[42] Xuelin Qian, Wenxuan Wang, Li Zhang, Fangrui Zhu, Yanwei Fu, Tao Xiang, Yu-Gang Jiang, and Xiangyang Xue. Long-term cloth-changing person re-identification. In *ACCV*, 2020. 3

[43] 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*, 2021. 3

[44] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, 2015. 2

[45] Zhiyin Shao, Xinyu Zhang, Changxing Ding, Jian Wang, and Jingdong Wang. Unified pre-training with pseudo texts for text-to-image person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11174–11184, 2023. 3

[46] Xiujun Shu, Ge Li, Xiao Wang, Weijian Ruan, and Qi Tian. Semantic-guided pixel sampling for cloth-changing person re-identification. *SPL*, 2021. 1

[47] Xiujun Shu, Xiao Wang, Xianghao Zang, Shiliang Zhang, Yuanqi Chen, Ge Li, and Qi Tian. Large-scale spatio-temporal person re-identification: Algorithms and benchmark. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(7):4390–4403, 2021. 3

[48] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *ECCV*, 2018. 7

[49] Shixiang Tang, Cheng Chen, Qingsong Xie, Meilin Chen, Yizhou Wang, Yuanzheng Ci, Lei Bai, Feng Zhu, Haiyang Yang, Li Yi, et al. Humanbench: Towards general human-centric perception with projector assisted pretraining. *arXiv preprint arXiv:2303.05675*, 2023. 8

[50] Fangbin Wan, Yang Wu, Xuelin Qian, Yixiong Chen, and Yanwei Fu. When person re-identification meets changing clothes. In *CVPR Workshops*, 2020. 3

[51] Zhikang Wang, Feng Zhu, Shixiang Tang, Rui Zhao, Lihuo He, and Jiangning Song. Feature erasing and diffusion network for occluded person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4754–4763, 2022. 1

[52] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *CVPR*, 2018. 1, 3

[53] Hu Xu, Gargi Ghosh, Po-Yao Huang, Prahal Arora, Masoumeh Aminzadeh, Christoph Feichtenhofer, Florian Metze, and Luke Zettlemoyer. Vlm: Task-agnostic video-language model pre-training for video understanding. *arXiv preprint arXiv:2105.09996*, 2021. 3

[54] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. Videoclip: Contrastive pre-training for zero-shot video-text understanding. *arXiv preprint arXiv:2109.14084*, 2021. 3

[55] Bin Yang, Mang Ye, Jun Chen, and Zesen Wu. Augmented dual-contrastive aggregation learning for unsupervised visible-infrared person re-identification. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 2843–2851, 2022. 1, 3

[56] Mouxing Yang, Zhenyu Huang, Peng Hu, Taihao Li, Jiancheng Lv, and Xi Peng. Learning with twin noisy labels for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14308–14317, 2022. 1, 3

[57] Qize Yang, Ancong Wu, and Wei-Shi Zheng. Person re-identification by contour sketch under moderate clothing change. *TPAMI*, 2019. 3

[58] Shuyu Yang, Yinan Zhou, Zhedong Zheng, Yaxiong Wang, Li Zhu, and Yujiao Wu. Towards unified text-based person retrieval: A large-scale multi-attribute and language search benchmark. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 4492–4501, 2023. 3

[59] Zhengwei Yang, Meng Lin, Xian Zhong, Yu Wu, and Zheng Wang. Good is bad: Causality inspired cloth-debiasing for cloth-changing person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1472–1481, 2023. 7

[60] Zhengwei Yang, Xian Zhong, Zhun Zhong, Hong Liu, Zheng Wang, and Shin'ichi Satoh. Win-win by competition: Auxiliary-free cloth-changing person re-identification. *IEEE Transactions on Image Processing*, 2023. 7

[61] Mang Ye, Weijian Ruan, Bo Du, and Mike Zheng Shou. Channel augmented joint learning for visible-infrared recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13567–13576, 2021. 7

[62] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. Deep learning for person re-identification: A survey and outlook. *TPAMI*, 2021. 3

[63] Zhou Yin, Wei-Shi Zheng, Ancong Wu, Hong-Xing Yu, Hai Wan, Xiaowei Guo, Feiyue Huang, and Jianhuang Lai. Adversarial attribute-image person re-identification. *arXiv preprint arXiv:1712.01493*, 2017. 1

[64] Shijie Yu, Shihua Li, Dapeng Chen, Rui Zhao, Junjie Yan, and Yu Qiao. Cocas: A large-scale clothes changing person dataset for re-identification. In *CVPR*, 2020. 1, 3, 7, 8

[65] Qiang Zhang, Changzhou Lai, Jianan Liu, Nianchang Huang, and Jungong Han. Fmcnet: Feature-level modality compensation for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7349–7358, 2022. 1, 3

[66] Renrui Zhang, Jiaming Han, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, Peng Gao, and Yu Qiao. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*, 2023. 4

[67] Yukang Zhang and Hanzi Wang. Diverse embedding expansion network and low-light cross-modality benchmark for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2153–2162, 2023. 3, 7

[68] Yukang Zhang, Yan Yan, Yang Lu, and Hanzi Wang. Towards a unified middle modality learning for visible-infrared person re-identification. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 788–796, 2021. 7

[69] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, Xin Jin, and Zhibo Chen. Relation-aware global attention for person re-identification. In *CVPR*, 2020. 7

[70] Bo Zhao, Shixiang Tang, Dapeng Chen, Hakan Bilen, and Rui Zhao. Continual representation learning for biometric identification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1198–1208, 2021. 1

[71] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jiahao Bu, and Qi Tian. Person re-identification meets image search. *arXiv preprint arXiv:1502.02171*, 2015. 3

[72] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *CVPR*, 2015. 1

[73] Liang Zheng, Yi Yang, and Alexander G Hauptmann. Person re-identification: Past, present and future. *arXiv preprint arXiv:1610.02984*, 2016. 1, 3

[74] Yi Zheng, Shixiang Tang, Guolong Teng, Yixiao Ge, Kaijian Liu, Jing Qin, Donglian Qi, and Dapeng Chen. Online pseudo label generation by hierarchical cluster dynamics for adaptive person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8371–8381, 2021. 1

[75] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *ICCV*, 2017. 1

[76] Zhedong Zheng, Liang Zheng, Michael Garrett, Yi Yang, Mingliang Xu, and Yi-Dong Shen. Dual-path convolutional image-text embeddings with instance loss. *TOMM*, 2020. 1, 3

[77] Xiao Zhou, Yujie Zhong, Zhen Cheng, Fan Liang, and Lin Ma. Adaptive sparse pairwise loss for object re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19691–19701, 2023. 3

[78] Kuan Zhu, Haiyun Guo, Tianyi Yan, Yousong Zhu, Jinqiao Wang, and Ming Tang. Pass: Part-aware self-supervised pre-training for person re-identification. In *European Conference on Computer Vision*, pages 198–214. Springer, 2022. 6, 8

[79] 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