# Co-Attention Aligned Mutual Cross-Attention for Cloth-Changing Person Re-Identification

Qizao Wang[1][0000−0003−2556−5529], Xuelin Qian*[2][0000−0001−8049−7288], Yanwei Fu[2][0000−0002−6595−6893], and Xiangyang Xue[1,2][0000−0002−4897−9209]

[1] School of Computer Science, Shanghai Key Lab of Intelligent Information Processing, Fudan University
qzwang22@m.fudan.edu.cn, xyxue@fudan.edu.cn
[2] School of Data Science, and MOE Frontiers Center for Brain Science, Shanghai Key Lab of Intelligent Information Processing, Fudan University
{xlqian,yanweifu}@fudan.edu.cn

**Abstract.** Person re-identification (Re-ID) has been widely studied and achieved significant progress. However, traditional person Re-ID methods primarily rely on cloth-related color appearance, which is unreliable under real-world scenarios when people change their clothes. Cloth-changing person Re-ID that takes this problem into account has received increasing attention recently, but it is more challenging to learn discriminative person identity features, since larger intra-class variation and smaller inter-class easily occur in the image feature space with clothing changes. Beyond appearance features, some known identity-related features can be implicitly encoded in images (*e.g.*, body shapes). In this paper, we first design a novel Shape Semantics Embedding (SSE) module to encode body shape semantic information, which is one of the essential clues to distinguish pedestrians when their clothes change. To better complement image features, we further propose a Co-attention Aligned Mutual Cross-attention (CAMC) framework. Different from previous attention-based fusion strategies, it first aligns features from multiple modalities, then effectively interacts and transfers identity-aware but cloth-irrelevant knowledge between the image space and the body shape space, resulting in a more robust feature representation. To the best of our knowledge, this is the first work to adopt Transformer to handle the multi-modal interaction for cloth-changing person Re-ID. Extensive experiments demonstrate the effectiveness of our proposed method and show the superior performance achieved on several cloth-changing person Re-ID benchmarks. Codes will be available at https://github.com/QizaoWang/CAMC-CCReID.

## 1 Introduction

Person re-identification (Re-ID) aims at identifying and associating the same person across different cameras, which has great potential applications in video

* corresponding author

surveillance, including suspect tracking, activity analysis, human-computer interaction, *etc.* Depending on application scenarios, existing person re-identification approaches can be broadly grouped into two categories, short-term and long-term. Short-term person Re-ID has been widely studied in the past decades, involved in multiple challenges and research directions, including occlusion [37, 30, 55] and infrared-visible modalities [53, 54], supervised [34, 26, 68] and unsupervised learning [51, 58, 21], representation [56, 50, 5] and metric learning [13, 7, 45]. However, all of these methods assume that the same person would always wear the same clothes, so the learned features may mostly rely on clothing appearances. On the contrary, long-term person Re-ID focuses more on the application in real-world scenarios, which takes into account practical problems, such as changing clothes and incremental identities. Among them, the cloth-changing problem has attracted more and more attention, which is also known as cloth-changing person Re-ID.

To deal with the challenge of clothing changes, it is important to use robust identity-related features. Many researchers naturally turn their attention to human body shape information, since the body shape of a person usually remains unchanged for a relatively long duration. However, it is extremely difficult to mine it from RGB color images. Consequently, most cloth-changing Re-ID methods draw support from other modalities, such as 2D human posture keypoints [39], contour sketches [57], gaits [22] and 3D shapes [6]. In this paper, we also target cloth-changing person Re-ID, but propose a novel Shape Semantics Embedding (SSE) module. It uses heatmaps of human postures to encode body shape semantic information, which is more lightweight and robust in long-term scenarios.

When integrating useful information from multiple modalities, one of the challenges is how to make them interact effectively. Unfortunately, in previous cloth-changing Re-ID methods, the interaction between appearance features and features extracted from other modalities is relatively simple. Intuitively, there is an inevitable gap in the representations of different modalities even for the same person, so a simple interaction between modalities could not make full of use of abundant multi-modal information. In recent years, Transformer [48] has been widely used, and many researches have shown its effectiveness in computer vision tasks, such as object detection [3, 70], and person re-identification [35, 30, 12]. The attention mechanism in Transformer can effectively capture the contextual semantic information of input sequences. Inspired by it, we propose a Co-Attention Aligned Mutual Cross-attention (CAMC) framework for cloth-changing person Re-ID.

More specifically, an appearance branch and a shape branch are first designed in our framework. The former uses a conventional backbone (*e.g.*, ResNet-50 [10]) to extract appearance features, and the latter is exactly our proposed SSE module for obtaining body shape information. To relieve the gap between both features from different modalities, we adopt element-wise attention for alignment. Subsequently, a symmetrical mutual cross-attention module is applied to effectively distill appearance and body shape information from each other. With our

proposed CAMC framework, appearance features are refined with the help of body shape semantic information which is cloth-irrelevant, while body shape semantic features are supplemented with aligned appearance features robust to clothing changes.

In summary, our contributions are listed as follows:

1. We propose a Shape Semantics Embedding module based on the self-attention mechanism to encode body shape semantics irrelevant to clothes, which is essential to identify a person when changing clothes.
2. We propose a novel mutual interaction module based on the cross-attention mechanism to interact appearance and body shape features effectively, resulting in a fused feature more robust to clothing changes. To mitigate the feature gap from different modalities and improve the efficiency and effectiveness of their feature interaction, we additionally introduce an element-wise co-attention alignment module for alignment.
3. To the best of our knowledge, it is the first work to adopt Transformer to handle multi-modal interaction for cloth-changing person Re-ID. Extensive experiments demonstrate the efficacy of our proposed model on several cloth-changing Re-ID benchmarks.

## 2 Related Work

### 2.1 Person Re-identification

Person Re-ID task aims at identifying a specific person across different cameras and locations. With the rise of deep learning, person Re-ID technology has made great progress and is widely used in smart cities, intelligent security, human-computer interaction, *etc.* Many works try to explore fine-grained pedestrian identity features via metric learning, for instance, hard triplet loss [7, 13] to encourage a closer feature distance among the same identity, and classification loss [41, 64, 66] to learn a high-level global feature from the whole input. There are also some other works dealing with spatial misalignment problems, such as occlusion [11, 65], variant camera views [33, 44], diverse poses [38, 28], different resolutions [27], and manifold domains [18, 23]. However, these models are well-trained based on the assumption that the same person has the same clothing in a short duration, which seriously hinders their applications in long-term real-world scenarios. In this paper, we focus on the more realistic long-term cloth-changing person Re-ID task and further explore a robust person identity feature extraction model to solve the problem of unreliable appearance information.

### 2.2 Cloth-changing Person Re-identication

To further improve the applicability and practicability of person Re-ID models in real-world scenarios, more and more researchers turn to studying the cloth-changing person Re-ID task, which targets to match the same person across different locations over a long duration and inevitably faces the cases of changing

clothes. In this situation, appearance/texture information can no longer be used as an accurate representation to distinguish different pedestrians, which makes it difficult for many previous methods to achieve satisfactory results. Thus, nowadays, many studies have tried to solve the problem via learning more stable biological representation, for example, Wan *et al.* [49] and Yu *et al.* [60] extract facial features to improve the person Re-ID accuracy; Yang *et al.* [57] utilize contour sketches to indicate discriminative characteristics; and Qian *et al.* [39] and Li *et al.* [29] use shape information to help feature learning. Different from existing works, we not only focus on extracting precise biological body shape semantic information, but also try to better align the two modality features of appearance and body shape, which can further boost the cloth-changing person Re-ID performance.

### 2.3 Transformer-based Person Re-identification

Transformer [48] has made great achievements in the field of natural language processing. Inspired by the self-attention mechanism, many researchers apply Transformers to computer vision tasks and find such Transformers can be as effective as CNNs over feature extraction. For example, Dosovitskiy *et al.* propose ViT [9] which processes images directly as sequences, Touvron *et al.* introduce a teacher-student strategy specific for Transformers to speed up the ViT training without using any large-scale pretraining data, and Carion *et al.* design DERT [3] performing cross-attention between the object query and the feature map to transform the detection task into a one-to-one matching problem. Since Transformer can capture long-distance dependency and help models pay attention to different parts of the human body, such as the head, shoulder, waist, and thigh, and obtain rich local relevant semantic information, Li *et al.* [30] and He *et al.* [12] adopt Transformer to solve the person partial-observation problem in occlusion person Re-ID task. In this paper, we take the advantage of the Transformer cross-attention mechanism to interact appearance and body shape semantic information, and generate a robust fused pedestrian feature under the cloth-changing scenario. Different from the latest work [2], which uses ViT as backbone only, we explore to effectively use Transformer to interact multiple modalities in cloth-changing person Re-ID.

## 3 Methodology

### 3.1 Overview

In this paper, we aim to address the problem of person Re-ID under the long-term cloth-changing setting, where the clothing appearance is unreliable and even would hinder the network to extract discriminative features. Considering that body shape is more robust against clothing changes, we draw support from heatmaps of human postures to encode body shape semantic information. Furthermore, we propose a Co-attention Aligned Mutual Cross-attention framework to effectively align and interact multi-modality features.

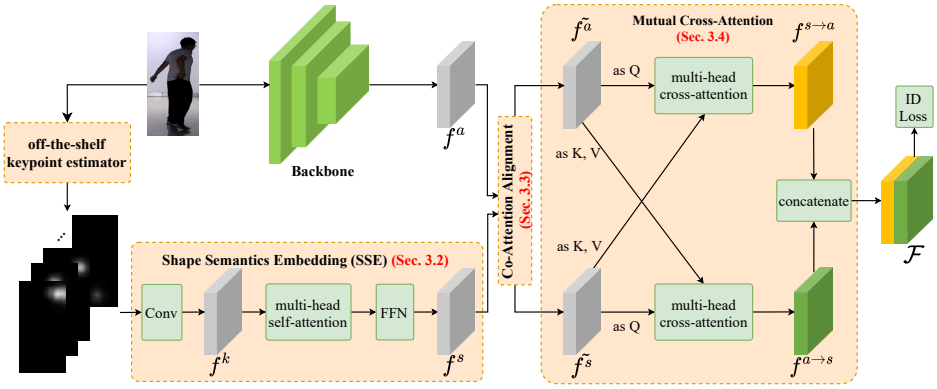


**Fig. 1.** Overview of our Co-attention Aligned Mutual Cross-attention (CAMC) framework. It consists of an appearance branch, and a body shape branch implemented by our proposed Shape Semantics Embedding (SSE) module. We propose a co-attention alignment module to align the two modalities. Then a symmetrical mutual cross-attention module is applied to effectively interact and fuse appearance and body shape semantic information, outputting a fused feature robust to clothing changes.

The overall framework is shown in Fig. 1, which consists of an appearance branch and a body shape branch. Concretely, the former is designed to extract appearance features from the given person image $x \in \mathbb{R}^{H\times W\times c}$. Following [46], we use ResNet-50 [10] as backbone, and the stride of the first convolution layer in *res4* block is set to 1 to increase the feature resolution. The output feature is flattened in the spatial dimensions to obtain the appearance feature sequence $f^a \in \mathbb{R}^{hw\times d}$, where $h$ and $w$ are the height and width of the feature map, $d$ denotes the feature dimension.

In the rest of this section, we first introduce how the body shape branch extracts rich semantic information about body shapes (see Sec. 3.2). Second, we propose a co-attention alignment module to align multi-modal information (see Sec. 3.3). Then, we elaborate on the mutual cross-attention module, which plays a key role in the multi-modal feature interaction (see Sec. 3.4). Lastly, we briefly describe the procedures of training and inference in Sec. 3.5.

### 3.2 Shape Semantics Embedding Module

When people change their clothes, although most appearance clues, such as the color of clothes, change significantly, their body shapes are relatively stable even for a long time. Therefore, it is useful to encode and utilize the body shape semantic information, which is more robust to clothing changes. To achieve such a goal, we propose a Shape Semantics Embedding (SSE) module to encode it from heatmaps of human postures. Our proposed SSE module especially leverages the self-attention mechanism to learn relations between different human posture keypoints. The intuition is that biological information is contained in

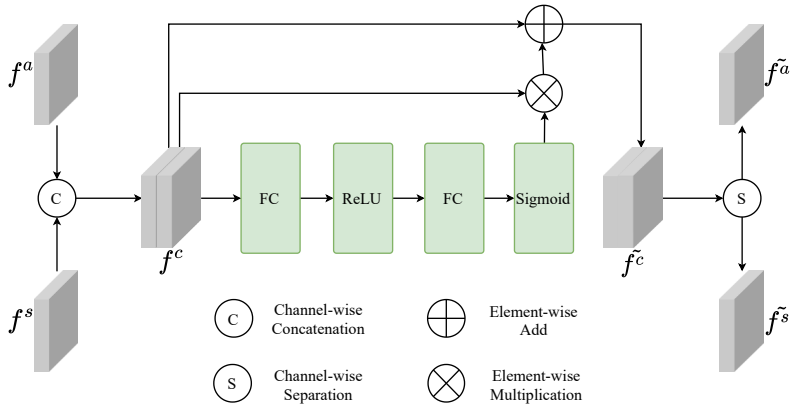


**Fig. 2.** Illustration of our proposed co-attention alignment module. By encoding the concatenated appearance and body shape semantic features, element-wise attention scores are obtained to effectively align features from different branches.

these relations (*e.g.*, hip to knee). Compared with encoding individual keypoint, such a design is more stable and robust, even if people change their postures.

As shown in Fig. 1, we employ an off-the-shelf estimator HRNet [43] to obtain heatmaps of human postures. Given an input image $x$, HRNet returns $K$ heatmaps of human postures, where each heatmap represents the location distribution of one posture keypoint. We regard $K$ as the feature dimension and apply a convolution layer to increase the dimension from $K$ to $d$, outputting a human posture feature $f^k \in \mathbb{R}^{hw \times d}$, where $h$ and $w$ represent the height and width of the feature map. To encode the relations between each pair of human posture parts, we feed $f^k$ into an encoder layer, which consists of one multi-head self-attention layer [48] with a skip connection and layer normalization [1]. A standard feed-forward network [48] is further attached to output the body shape semantic feature $f^s \in \mathbb{R}^{hw \times d}$. More details on the structure design are discussed in the supplemental material.

Benefiting from the ability to effectively capture the long-distance and short-distance semantic information of inputs, the self-attention mechanism can encode the correlations between different parts of human postures, which present the body shape semantics. Meanwhile, the multi-head mechanism enables different heads to effectively focus on all kinds of semantic details.

### 3.3 Co-Attention Alignment Module

Intuitively, the appearance feature $f^a$ and the body shape feature $f^s$ are from two different modalities, so they may not correspond, containing mismatched redundancy and interference information. To effectively match and integrate their useful information, we propose a co-attention alignment module, as shown in Fig. 2, to align both features before further interaction and fusion.

Specifically, the input of the co-attention alignment module is the concatenated two branch features $f^c = [f^a; f^s] \in \mathbb{R}^{hw \times 2d}$, where $[* \ ; \ *]$ means the concatenation operation. Then it further goes through two fully-connected (FC) layers. Similar to [16], the first FC layer compresses the feature dimension to a quarter and the second one decodes it back to the original dimension, not only reducing the number of parameters, but also achieving the information bottleneck effect. After that, a *sigmoid* function is attached to produce element-wise attention scores for feature alignment. For training stability, we further introduce a skip connection from the input to the output. The overall formulation can be expressed as,

$$s = \sigma(\phi(f^c W_1 + b_1)W_2 + b_2) \tag{1}$$

$$[\tilde{f}^a; \tilde{f}^s] = \tilde{f}^c \ ; \quad \tilde{f}^c = s \otimes f^c + f^c \tag{2}$$

where $W_1 \in \mathbb{R}^{2d \times (d/2)}$, $W_2 \in \mathbb{R}^{(d/2) \times 2d}$, $b_1 \in \mathbb{R}^{1 \times (d/2)}$ and $b_2 \in \mathbb{R}^{1 \times 2d}$ are weights and biases of two fully-connected layers; $\phi$ is the $ReLU$ activation function and $\sigma$ denotes the *sigmoid* function; $\otimes$ means the element-wise multiplication. We divide the aligned features $\tilde{f}^c \in \mathbb{R}^{hw \times 2d}$ in half, to separately get the aligned appearance feature sequence $\tilde{f}^a \in \mathbb{R}^{hw \times d}$ and the aligned body shape semantic feature sequence $\tilde{f}^s \in \mathbb{R}^{hw \times d}$.

### 3.4 Mutual Cross-Attention Module

After successfully aligning appearance and body shape semantic features, we introduce a cross-attention module to realize the information interaction between the two modalities. As shown in Fig. 1, this module is performed mutually and symmetrically, that is, the body shape features are integrated into the appearance features and vice versa.

Take one side as an example, we first apply the multi-head cross-attention mechanism to do interaction by computing the dot-product similarity between the appearance feature $\tilde{f}^a$ and the body shape feature $\tilde{f}^s$. The similarity is scaled by $\sqrt{d}$ and normalized by a softmax function. Subsequently, the result is regarded as an attention weight to perform a weighted sum of $\tilde{f}^s$. In this way, for each appearance feature $\tilde{f}^a_i \in \mathbb{R}^{1 \times d}$, where $i \in [1, hw]$, we can find body shape semantic information with similar responses in $\tilde{f}^s$, and integrate it effectively. In other words, the appearance features $\tilde{f}^a$ are well refined with the help of the cloth-irrelevant body shape semantic information $\tilde{f}^s$. The formulation is expressed as,

$$Q = \tilde{f}^a \ ; \quad K = \tilde{f}^s \ ; \quad V = \tilde{f}^s \tag{3}$$

$$f^{s \to a} = \psi\left(\text{softmax}\left(\frac{QK^{\intercal}}{\sqrt{d}}\right)V + \tilde{f}^a\right) \tag{4}$$

where $f^{s \to a} \in \mathbb{R}^{hw \times d}$ and $\psi$ denotes the layer normalization [1]. It is similar to the standard multi-head cross-attention module in Transformer [48], but there

is no necessary for linear projection for query, key, and value space, thanks to our proposed co-attention alignment module.

For the other side of the interaction, we adopt the same formulation but set $Q = \tilde{f}^s$, $K = \tilde{f}^a$, $V = \tilde{f}^a$, to integrate the matched appearance features into the body shape semantic features. Finally, we get $f^{a\to s} \in \mathbb{R}^{hw\times d}$, which represents body shape semantic features that are supplemented with aligned appearance features robust to clothing changes.

### 3.5 Training and Inference

Thanks to the symmetrical mutual cross-attention interaction between the aligned features from the two branches, we take full use of appearance and body shape semantic information. We concatenate the two features after interaction together to obtain a more robust and discriminative feature, which can be formulated as follows:

$$\mathcal{F} = [f^{s\to a}; f^{a\to s}] \in \mathbb{R}^{hw\times 2d} \tag{5}$$

For training, we adopt a linear classifier with the input of the fused feature $\mathcal{F}$, and optimize the model by cross-entropy loss with label smoothing. For inference, we directly use $\mathcal{F}$ as final features to compute the cosine distance between two person images for retrieval.

## 4 Experiment

### 4.1 Experimental Setup

**Datasets.** We mainly evaluate our method on two widely used long-term cloth-changing person Re-ID datasets: Celeb-reID [20] and LTCC [39]. **Celeb-reID** is acquired from the Internet using street snapshots of celebrities, which contains 34,186 images of 1,052 identities. Specifically, more than 70% images of each person show different clothes on average. **LTCC** is an indoor cloth-changing person Re-ID dataset, which has 17,138 images of 152 identities with 478 different outfits captured from 12 camera views. LTCC is challenging as it contains diverse human poses, large changes of illumination, and large variations of occlusion. To better illustrate our model efficacy on the general person Re-ID task, we also evaluate our method on **Market-1501** [63], which is a benchmark dataset for the standard person Re-ID without clothing changes.

**Implementation details.** Our method is implemented on the Pytorch framework. We adopt ResNet-50 [10] initialized by ImageNet [8] as backbone to extract person appearance features. The input images are resized to $256 \times 128$. For data augmentation, color jitter, random horizontal flipping, padding, random cropping, and random erasing [67] are used. We use Adam optimizer [24] for 150 epochs, with the warmup strategy that linearly increases the learning rate from $3 \times 10^{-5}$ to $3 \times 10^{-4}$ in the first 10 epochs. Then decrease the learning rate by a factor of 10 at epoch 40 and 80. The batch size is set to 64 for Celeb-reID and Market-1501, and 32 for LTCC, with 4 images per ID. To get the heatmaps

**Table 1.** Comparison of our method with the state-of-the-art methods on Celeb-reID. The best results are shown in bold.

| Methods | Rank-1 | Rank-5 | mAP |
|---|---|---|---|
| ResNet-Mid [59] | 43.3 | 54.6 | 5.8 |
| Two-Stream [66] | 36.3 | 54.5 | 7.8 |
| MLFN [4] | 41.4 | 54.7 | 6.0 |
| HACNN [26] | 47.6 | 63.3 | 9.5 |
| Part-Bilinear [42] | 19.4 | 40.6 | 6.4 |
| PCB [46] | 37.1 | 57.0 | 8.2 |
| MGN [52] | 49.0 | 64.9 | 10.8 |
| ReIDCaps [20] | 51.2 | 65.4 | 9.8 |
| CESD [39] | 50.9 | 66.3 | 9.8 |
| RCSANet [19] | 54.9 | - | 11.0 |
| Baseline (ResNet-50) | 52.9 | 66.2 | 9.9 |
| Ours | **57.5** | **71.5** | **12.3** |

of human postures, we employ HRNet [43] pre-trained on COCO dataset [32], where the number of heatmaps is 17. We merge the 5 heatmaps corresponding to the nose, ears, and eyes as "face", resulting in 13 heatmaps. We simply freeze all weights of HRNet during training.

**Evaluation metrics.** For evaluation, we adopt standard metrics as in most person Re-ID literature, namely Cumulative Matching Characteristic (CMC) curves and mean average precision (mAP). To make a fair comparison with the existing research works, for LTCC, we evaluate our method under both the standard setting and the cloth-changing setting. Specifically, for the standard setting, images in the testing set with the same identity and the same camera view are discarded when computing evaluation scores. In other words, there are both cloth-consistent and cloth-changing samples in the testing set. For the cloth-changing setting, images with the same identity, camera view, and clothes are discarded during testing, so there are only cloth-changing samples in the testing set.

### 4.2 Quantitative Results

**Performance on the Celeb-reID dataset.** We evaluate our proposed method on Celeb-reID and compare it with other state-of-the-art competitors. Results are shown in Table 1. Among them, ReIDCaps [20], CESD [39] and RCSANet [19] are specially designed for the cloth-changing person Re-ID problem. For a fair comparison, the results of ReIDCaps [20] and RCSANet [19] are achieved without applying the fine-grained body parts learning strategy. The results of ReIDCaps [20] are copied from the original paper and it uses deeper DenseNet-121 [17] as the backbone. Our method outperforms all compared methods on the challenging cloth-changing dataset Celeb-reID which contains great clothing variations. Our method outperforms the state-of-the-art method RCSANet [19]

**Table 2.** Comparison of our method with the state-of-the-art methods on LTCC. The best results of the state-of-the-art method and our method are shown in bold. "Standard" and "Cloth-Changing" mean the standard setting and the cloth-changing setting, respectively.

| Methods | Cloth-Changing | | Standard | |
|---|---|---|---|---|
| | Rank-1 | mAP | Rank-1 | mAP |
| LOMO [31] + KISSME [25] | 10.75 | 5.25 | 26.57 | 9.11 |
| LOMO [31] + XQDA [31] | 10.95 | 6.2 | 25.35 | 9.54 |
| PCB [46] | 23.52 | 10.03 | 61.86 | 27.52 |
| HACNN [26] | 21.59 | 9.25 | 60.24 | 26.71 |
| RGA-SC [62] | 31.4 | 14.0 | 65.0 | 27.5] |
| ISP [69] | 27.8 | 11.9 | 66.3 | 29.6 |
| GI-ReID [22] | 23.7 | 10.4 | 63.2 | 29.4 |
| CESD [39] | 25.15 | 12.40 | 71.39 | 34.41 |
| Chen *et al.* [6] | 31.2 | 14.8 | - | - |
| FSAM [14] | **38.5** | **16.2** | **73.2** | **35.4** |
| Baseline (ResNet-50) | 31.89 | 13.07 | 69.17 | 33.16 |
| Ours | **35.97** | **15.43** | **73.23** | **35.31** |

by 2.6% in Rank-1 accuracy. The great improvement of our method compared with our baseline model (ResNet-50) also demonstrates that our method can help tackle the cloth-changing challenge of person Re-ID.

**Performance on the LTCC dataset.** We also evaluate our proposed method on LTCC and compare it with several competitors. In Table 2, competitors include methods based on hand-crafted feature representations, deep learning baselines, and methods specially designed for cloth-changing person Re-ID. All state-of-the-art standard person Re-ID methods achieve relatively inferior performance, because they do not take the clothing changes into account. To reduce the interference of clothes, some cloth-changing person Re-ID methods use information from different modalities. For example, FSAM [14] integrates three modalities and fine-tunes the parsing network while training, while our method only uses an off-the-shelf human posture keypoints extractor. Results show that our method achieves comparable results with the state-of-the-art cloth-changing person Re-ID methods.

**Performance on the Market-1501 dataset.** To further show the feasibility of our method for the cases without clothing changes in the short term, we additionally evaluate our method on the standard benchmark person Re-ID dataset Market-1501. As shown in Table 3, our method is comparable with the state-of-the-art methods on Market-1501. Specifically, our method still achieves improvement compared with the baseline model (ResNet-50), which shows that our method can take advantage of the body shape information to extract more discriminative person identity features.

**Table 3.** Comparison of our method with state-of-the-art methods on Market-1501.

| Methods | Rank-1 | mAP |
|---|---|---|
| PCB [46] | 93.8 | 81.6 |
| IANet [15] | 94.9 | 83.1 |
| AANet [47] | 93.9 | 83.4 |
| DSA-reID [61] | 95.7 | 87.6 |
| RGA-SC [62] | 96.1 | 88.4 |
| ISP [69] | 95.3 | 88.6 |
| Baseline (ResNet-50) | 93.1 | 82.9 |
| Ours | 94.0 | 84.6 |

**Table 4.** Ablation study on the Celeb-reID dataset. "S→A" denotes the one-way cross-attention interaction from the body shape branch to the appearance branch, while "A→S" denotes the one-way cross-attention interaction from the appearance branch to the body shape branch.

| Methods | SSE | Co-Attention | S→A | A→S | Rank-1 | mAP |
|---|---|---|---|---|---|---|
| 1 (Baseline) | | | | | 52.86 | 9.92 |
| 2 | ✓ | | | | 52.59 | 9.97 |
| 3 | ✓ | ✓ | | | 53.23 | 10.19 |
| 4 | ✓ | | ✓ | | 54.24 | 10.51 |
| 5 | ✓ | | | ✓ | 55.85 | 11.37 |
| 6 | ✓ | | ✓ | ✓ | 55.92 | 11.27 |
| 7 | ✓ | ✓ | ✓ | | 53.87 | 10.39 |
| 8 | ✓ | ✓ | | ✓ | 57.17 | 12.09 |
| 9 (Ours) | ✓ | ✓ | ✓ | ✓ | **57.47** | **12.27** |

### 4.3 Ablation Study

To verify the effectiveness of our method, detailed ablation experiments are carried out on each proposed module, on the large-scale long-term cloth-changing person Re-ID dataset Celeb-reID. Results are shown in Table 4.

**The effectiveness of SSE and mutual cross-attention.** Although body shape is more robust against clothing changes than color appearance, intuitively, we cannot distinguish a person only by his/her body shape. Experiments also demonstrate that the performance is quite low if we only use the body shape branch. The results of method 2 in Table 4 show that if we directly concatenate appearance features with body shape semantic features extracted by the SSE module, the performance is close to baseline. It shows the performance improvement is gained from our well-designed mutual cross-attention strategy, rather than just the extra introduction of the body shape branch. By comparing methods 2 and 6 in Table 4, we can see that our proposed mutual cross-attention strategy improves Rank-1 by 3.33%, and mAP by 1.30%. It also indicates that the SSE module has encoded useful body shape semantic information.

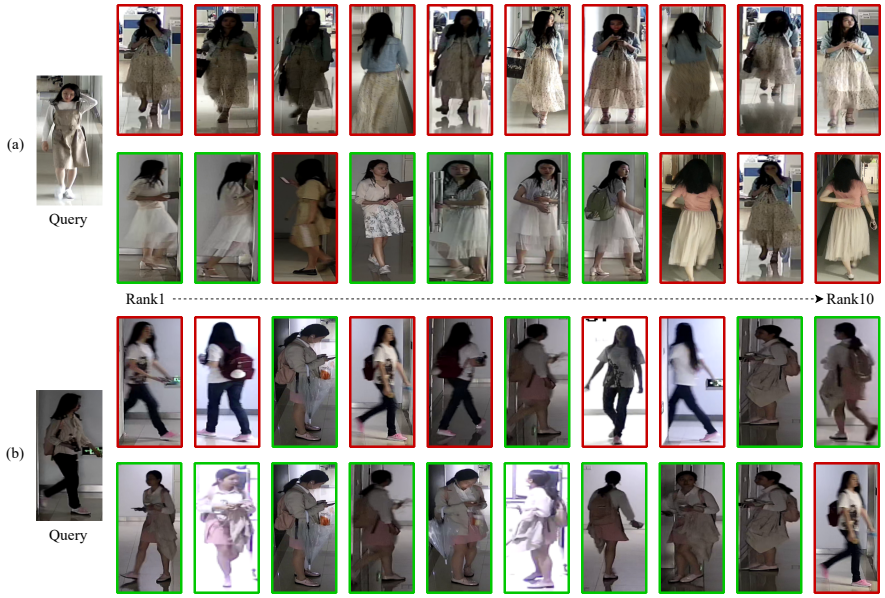


**Fig. 3.** Visualization of retrieval results. The left side of (a) and (b) is the input query image. For the right side, the first and the second row are the ordered matching results obtained by using the benchmark network ResNet-50 and our proposed network, respectively. Images with green borders and red borders indicate correct and error matching results, respectively. Best viewed in color and zoomed in.

**Design effectiveness of mutual cross-attention.** Aiming for adequate interaction and information fusion between the two branches, we propose a symmetrical mutual cross-attention module. As shown in Table 4, compared with baseline, when we only apply the one-way cross-attention interaction either "S→A" or "A→S", the performance is improved. However, due to the unidirectional nature of information interaction, the network still cannot make full use of the information between the two modalities. When we apply the proposed mutual cross-attention interaction strategy, much greater improvement is achieved, which validates the effectiveness of our mutual cross-attention design.

It is worth noting that, as the results of methods 4 and 7 in Table 4 show, if only use the one-way cross-attention interaction, applying the co-attention alignment mechanism may not improve the performance effectively. Therefore, our proposed mutual cross-attention strategy is more stable and conducive to multi-modal feature fusion.

**The effectiveness of the proposed co-attention alignment module.** As shown in Table 4, even without the mutual cross-attention module, compared with baseline, the co-attention alignment module can still improve the performance. Together with our proposed mutual cross-attention module, the perfor-

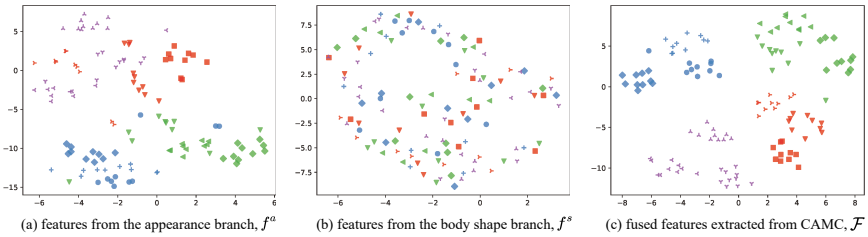


**Fig. 4.** t-SNE visualizations of features from the appearance branch and the body shape branch, as well as ones extracted from CAMC. Samples are randomly selected from the testing set of the LTCC dataset. Each color represents an identity, and different symbols indicate different clothes. Best viewed in color and zoomed in.

mance can be further improved. It is worth noting that, as the results of methods $4 \sim 6$ in Table 4 shown, when we discard our proposed co-attention alignment module, the mutual cross-attention strategy may not improve the performance effectively compared with the one-way cross-attention. The results confirm the necessity and effectiveness of our co-attention alignment operation. As the results of methods $7 \sim 9$ in Table 4 shown, when we apply our proposed co-attention aligned mutual cross-attention mechanism, the best results are achieved. It indicates features from various modalities, that are aligned with each other well, can interact and fuse more effectively, and better help the network to extract more discriminative and cloth-irrelevant identity features.

**Visualization of retrieval results.** With the introduction of the body shape semantic information and our proposed modality-aligned mutual fusion strategy, our method can help meet the challenge of changing clothes. To intuitively demonstrate this conclusion, we visualize the top 10 ranked retrieval results of the baseline network ResNet-50 and our proposed network under the cloth-changing setting on LTCC.

As shown in Fig. 3, our proposed network can better recognize the same person with different clothes. For example, in the second row of Fig. 3 (a), the top retrieval results have the same identity as the input query person, but with different clothes. However, for the matching results of ResNet-50, that is, the first row of Fig. 3 (a), the matching images have similar clothing textures to the input query image, but with different identities. For another example, we can see in Fig. 3 (b), the retrieved persons in the first row wear clothes with similar colors, resulting in some matching errors. However, benefiting from paying more attention to body shape information rather than volatile color appearance, our method can still correctly identify pedestrians even if they change clothes.

**Visualization of features.** To verify our motivation and show the effectiveness of our proposed method, we use t-SNE [36] to visualize the learned features.

As shown in Fig. 4, features from the appearance branch are relatively more chaotic than ones extracted from CAMC, indicating that different persons are misidentified under the influence of similar clothes. We can observe that although features from the body shape branch themselves are randomly distributed, our proposed CAMC framework can make use of them to obtain more discriminative fused features. More discussions and analyses are provided in the supplemental material.

## 5 Conclusion

In this paper, we study the more realistic and challenging long-term cloth-changing person Re-ID problem and propose a unified framework adopting Transformer to handle multiple modalities for the first time. Especially, with our proposed Shape Semantics Embedding (SSE) module, we can extract body shape semantic features, which are robust against clothing changes in the long term. To further integrate and make full use of the body shape semantic information, we propose a Co-attention Aligned Mutual Cross-attention (CAMC) framework and effectively fuse multiple modalities. As a result, features encoding useful appearance and body shape semantic information are distilled to an identity-related and discriminative feature, that is more robust to clothing changes. The effectiveness of our proposed method is validated through extensive experiments on several datasets.

**Broader Impact.** Our proposed CAMC framework can be easily used in existing person Re-ID methods to make long-term person Re-ID technology more practicable in intelligent video monitoring systems, and hopefully inspire more valuable and innovative studies. However, in reality, person Re-ID systems typically use unauthorized surveillance data, which may cause privacy breaches. As a result, governments and officials must take action to govern the use of person Re-ID data and technology, and researchers should avoid using datasets that may raise ethical concerns. For example, the dataset DukeMTMC [40] should no longer be used after it was shut down for violating data collection restrictions. It is worth noting that, all datasets used in our paper are publicly available and involve no ethical issues.

**Acknowledgements.** This work is supported by China Postdoctoral Science Foundation (2022M710746), the Science and Technology Major Project of Commission of Science and Technology of Shanghai (No.21XD1402500), NSFC Project (62176061).

## References

1. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization. arXiv preprint arXiv:1607.06450 (2016)

2. Bansal, V., Foresti, G.L., Martinel, N.: Cloth-changing person re-identification with self-attention. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 602–610 (2022)
3. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European conference on computer vision. pp. 213–229. Springer (2020)
4. Chang, X., Hospedales, T.M., Xiang, T.: Multi-level factorisation net for person re-identification. In: CVPR. vol. 1, p. 2 (2018)
5. Chen, B., Deng, W., Hu, J.: Mixed high-order attention network for person re-identification. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 371–381 (2019)
6. Chen, J., Jiang, X., Wang, F., Zhang, J., Zheng, F., Sun, X., Zheng, W.S.: Learning 3d shape feature for texture-insensitive person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8146–8155 (2021)
7. Chen, W., Chen, X., Zhang, J., Huang, K.: Beyond triplet loss: a deep quadruplet network for person re-identification. In: Proc. CVPR. vol. 2 (2017)
8. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. pp. 248–255. IEEE (2009)
9. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2015)
11. He, L., Liang, J., Li, H., Sun, Z.: Deep spatial feature reconstruction for partial person re-identification: Alignment-free approach. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7073–7082 (2018)
12. He, S., Luo, H., Wang, P., Wang, F., Li, H., Jiang, W.: Transreid: Transformer-based object re-identification. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 15013–15022 (2021)
13. Hermans, A., Beyer, L., Leibe, B.: In defense of the triplet loss for person re-identification. arXiv preprint arXiv:1703.07737 (2017)
14. Hong, P., Wu, T., Wu, A., Han, X., Zheng, W.S.: Fine-grained shape-appearance mutual learning for cloth-changing person re-identification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10513–10522 (2021)
15. Hou, R., Ma, B., Chang, H., Gu, X., Shan, S., Chen, X.: Interaction-and-aggregation network for person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9317–9326 (2019)
16. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7132–7141 (2018)
17. Huang, G., Liu, Z., Weinberger, K.Q., van der Maaten, L.: Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. vol. 1, p. 3 (2017)
18. Huang, Y., Wu, Q., Xu, J., Zhong, Y.: Sbsgan: Suppression of inter-domain background shift for person re-identification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9527–9536 (2019)

19. Huang, Y., Wu, Q., Xu, J., Zhong, Y., Zhang, Z.: Clothing status awareness for long-term person re-identification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11895–11904 (2021)
20. Huang, Y., Xu, J., Wu, Q., Zhong, Y., Zhang, P., Zhang, Z.: Beyond scalar neuron: Adopting vector-neuron capsules for long-term person re-identification. IEEE Transactions on Circuits and Systems for Video Technology (2019)
21. Isobe, T., Li, D., Tian, L., Chen, W., Shan, Y., Wang, S.: Towards discriminative representation learning for unsupervised person re-identification. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 8526–8536 (2021)
22. Jin, X., He, T., Zheng, K., Yin, Z., Shen, X., Huang, Z., Feng, R., Huang, J., Chen, Z., Hua, X.S.: Cloth-changing person re-identification from a single image with gait prediction and regularization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14278–14287 (2022)
23. Jin, X., Lan, C., Zeng, W., Chen, Z., Zhang, L.: Style normalization and restitution for generalizable person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3143–3152 (2020)
24. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
25. Koestinger, M., Hirzer, M., Wohlhart, P., Roth, P.M., Bischof, H.: Large scale metric learning from equivalence constraints. In: CVPR (2012)
26. Li, W., Zhu, X., Gong, S.: Harmonious attention network for person re-identification. In: CVPR. vol. 1, p. 2 (2018)
27. Li, Y.J., Chen, Y.C., Lin, Y.Y., Du, X., Wang, Y.C.F.: Recover and identify: A generative dual model for cross-resolution person re-identification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8090–8099 (2019)
28. Li, Y.J., Lin, C.S., Lin, Y.B., Wang, Y.C.F.: Cross-dataset person re-identification via unsupervised pose disentanglement and adaptation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 7919–7929 (2019)
29. Li, Y.J., Luo, Z., Weng, X., Kitani, K.M.: Learning shape representations for clothing variations in person re-identification. arXiv preprint arXiv:2003.07340 (2020)
30. Li, Y., He, J., Zhang, T., Liu, X., Zhang, Y., Wu, F.: Diverse part discovery: Occluded person re-identification with part-aware transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2898–2907 (2021)
31. Liao, S., Hu, Y., Zhu, X., Li., S.Z.: Person re-identification by local maximal occurrence representation and metric learning. In: CVPR (2015)
32. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
33. Liu, F., Zhang, L.: View confusion feature learning for person re-identification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6639–6648 (2019)
34. Luo, H., Gu, Y., Liao, X., Lai, S., Jiang, W.: Bag of tricks and a strong baseline for deep person re-identification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. pp. 0–0 (2019)
35. Luo, H., Jiang, W., Fan, X., Zhang, C.: Stnreid: Deep convolutional networks with pairwise spatial transformer networks for partial person re-identification. IEEE Transactions on Multimedia **22**(11), 2905–2913 (2020)
36. Maaten, L.v.d., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(Nov), 2579–2605 (2008)

37. Miao, J., Wu, Y., Liu, P., Ding, Y., Yang, Y.: Pose-guided feature alignment for occluded person re-identification. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 542–551 (2019)
38. Qian, X., Fu, Y., Wang, W., Xiang, T., Wu, Y., Jiang, Y.G., Xue, X.: Pose-normalized image generation for person re-identification. ECCV (2018)
39. Qian, X., Wang, W., Zhang, L., Zhu, F., Fu, Y., Xiang, T., Jiang, Y.G., Xue, X.: Long-term cloth-changing person re-identification. In: Proceedings of the Asian Conference on Computer Vision (2020)
40. Ristani, E., Solera, F., Zou, R., Cucchiara, R., Tomasi, C.: Performance measures and a data set for multi-target, multi-camera tracking. In: European Conference on Computer Vision workshop on Benchmarking Multi-Target Tracking (2016)
41. Shen, Y., Li, H., Yi, S., Chen, D., Wang, X.: Person re-identification with deep similarity-guided graph neural network. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 486–504 (2018)
42. Suh, Y., Wang, J., Tang, S., Mei, T., Mu Lee, K.: Part-aligned bilinear representations for person re-identification. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 402–419 (2018)
43. Sun, K., Xiao, B., Liu, D., Wang, J.: Deep high-resolution representation learning for human pose estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5693–5703 (2019)
44. Sun, X., Zheng, L.: Dissecting person re-identification from the viewpoint of viewpoint. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 608–617 (2019)
45. Sun, Y., Cheng, C., Zhang, Y., Zhang, C., Zheng, L., Wang, Z., Wei, Y.: Circle loss: A unified perspective of pair similarity optimization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6398–6407 (2020)
46. Sun, Y., Zheng, L., Yang, Y., Tian, Q., Wang, S.: Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 480–496 (2018)
47. Tay, C.P., Roy, S., Yap, K.H.: Aanet: Attribute attention network for person re-identifications. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7134–7143 (2019)
48. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems. pp. 5998–6008 (2017)
49. Wan, F., Wu, Y., Qian, X., Chen, Y., Fu, Y.: When person re-identification meets changing clothes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 830–831 (2020)
50. Wang, C., Zhang, Q., Huang, C., Liu, W., Wang, X.: Mancs: A multi-task attentional network with curriculum sampling for person re-identification. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 365–381 (2018)
51. Wang, D., Zhang, S.: Unsupervised person re-identification via multi-label classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10981–10990 (2020)
52. Wang, G., Yuan, Y., Chen, X., Li, J., Zhou, X.: Learning Discriminative Features with Multiple Granularities for Person Re-Identification. ArXiv e-prints (Apr 2018)

53. Wang, Z., Wang, Z., Zheng, Y., Chuang, Y.Y., Satoh, S.: Learning to reduce dual-level discrepancy for infrared-visible person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 618–626 (2019)
54. Wu, Q., Dai, P., Chen, J., Lin, C.W., Wu, Y., Huang, F., Zhong, B., Ji, R.: Discover cross-modality nuances for visible-infrared person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4330–4339 (2021)
55. Yan, C., Pang, G., Jiao, J., Bai, X., Feng, X., Shen, C.: Occluded person re-identification with single-scale global representations. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11875–11884 (2021)
56. Yang, F., Yan, K., Lu, S., Jia, H., Xie, X., Gao, W.: Attention driven person re-identification. Pattern Recognition **86**, 143–155 (2019)
57. Yang, Q., Wu, A., Zheng, W.S.: Person re-identification by contour sketch under moderate clothing change. IEEE Transactions on Pattern Analysis and Machine Intelligence (2019)
58. Yu, H.X., Zheng, W.S., Wu, A., Guo, X., Gong, S., Lai, J.H.: Unsupervised person re-identification by soft multilabel learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2148–2157 (2019)
59. Yu, Q., Chang, X., Song, Y.Z., Xiang, T., Hospedales, T.M.: The devil is in the middle: Exploiting mid-level representations for cross-domain instance matching. arXiv preprint arXiv:1711.08106 (2017)
60. Yu, S., Li, S., Chen, D., Zhao, R., Yan, J., Qiao, Y.: Cocas: A large-scale clothes changing person dataset for re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3400–3409 (2020)
61. Zhang, Z., Lan, C., Zeng, W., Chen, Z.: Densely semantically aligned person re-identification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 667–676 (2019)
62. Zhang, Z., Lan, C., Zeng, W., Jin, X., Chen, Z.: Relation-aware global attention for person re-identification. In: Proceedings of the ieee/cvf conference on computer vision and pattern recognition. pp. 3186–3195 (2020)
63. Zheng, L., Shen, L., Tian, L., S.Wang, J.Wang, Tian, Q.: Scalable person re-identification: A benchmark. In: ICCV (2015)
64. Zheng, L., Zhang, H., Sun, S., Chandraker, M., Tian, Q.: Person re-identification in the wild. arXiv preprint arXiv:1604.02531 (2016)
65. Zheng, W.S., Li, X., Xiang, T., Liao, S., Lai, J., Gong, S.: Partial person re-identification. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4678–4686 (2015)
66. Zheng, Z., Zheng, L., Yang, Y.: A discriminatively learned cnn embedding for person reidentification. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) **14**(1), 13 (2017)
67. Zhong, Z., Zheng, L., Kang, G., Li, S., Yang, Y.: Random erasing data augmentation. arXiv preprint arXiv:1708.04896 (2017)
68. Zhou, K., Yang, Y., Cavallaro, A., Xiang, T.: Omni-scale feature learning for person re-identification. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3702–3712 (2019)
69. Zhu, K., Guo, H., Liu, Z., Tang, M., Wang, J.: Identity-guided human semantic parsing for person re-identification. In: European Conference on Computer Vision. pp. 346–363. Springer (2020)

70. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable detr: Deformable transformers for end-to-end object detection. arXiv preprint arXiv:2010.04159 (2020)