Identity-Guided Human Semantic Parsing for Person Re-Identification

Kuan Zhu[1,2], Haiyun Guo[1], Zhiwei Liu[1,2], Ming Tang[1,3], and Jinqiao Wang[1,2]

[1] National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
[3] Shenzhen Infinova Limited, Shenzhen, China
{kuan.zhu, haiyun.guo, zhiwei.liu, tangm, jqwang}@nlpr.ia.ac.cn

Abstract. Existing alignment-based methods have to employ the pre-trained human parsing models to achieve the pixel-level alignment, and cannot identify the personal belongings (e.g., backpacks and reticule) which are crucial to person re-ID. In this paper, we propose the identity-guided human semantic parsing approach (ISP) to locate both the human body parts and personal belongings at pixel-level for aligned person re-ID only with person identity labels. We design the cascaded clustering on feature maps to generate the pseudo-labels of human parts. Specifically, for the pixels of all images of a person, we first group them to foreground or background and then group the foreground pixels to human parts. The cluster assignments are subsequently used as pseudo-labels of human parts to supervise the part estimation and ISP iteratively learns the feature maps and groups them. Finally, local features of both human body parts and personal belongings are obtained according to the self-learned part estimation, and only features of visible parts are utilized for the retrieval. Extensive experiments on three widely used datasets validate the superiority of ISP over lots of state-of-the-art methods. Our code is available at https://github.com/CASIA-IVA-Lab/ISP-reID.

Keywords: person re-ID, weakly-supervised human parsing, aligned representation learning

1 Introduction

Person re-identification (re-ID), which aims to associate the person images captured by different cameras from various viewpoints, has attracted increasing attention from both the academia and the industry. However, the task of person re-ID is inherently challenging because of the ubiquitous misalignment issue, which is commonly caused by part occlusions, inaccurate person detection, human pose variations or camera viewpoints changing. All these factors can significantly change the visual appearance of a person in images and greatly increase the difficulty of this retrieval problem.

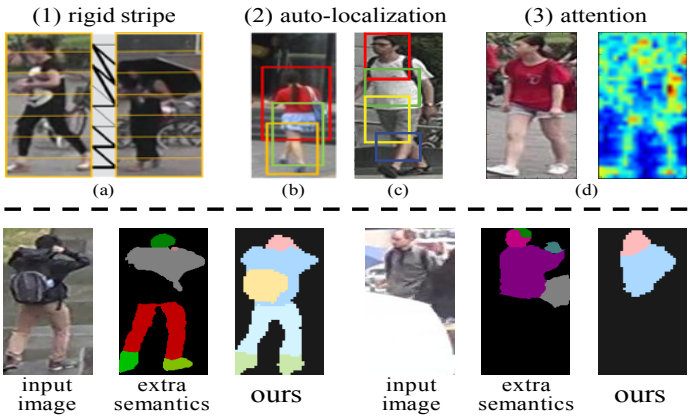

Fig. 1. The alignment-based methods. From (a) to (d): AlignedReID [52], MSCAN [20], DPL [51], MHN [3]. The extra semantic in the second row is predicted by the pre-trained parsing model [21], which exclude the personal belongings and are error-prone when one person is occluded by another. Our method is the first extra semantic free method which can locate both the human body parts and personal belongings at pixel-level, and explicitly identify the visible parts in an occluded image

In recent years, plenty of efforts have been made to alleviate the misalignment problem. The extra semantic free methods try to address the misalignment issue through a self-learned style. However, they can only achieve coarse alignment at region-level. These methods could be roughly summarized to the following streams: (1) The rigid stripe based methods, which directly partition the person image into fixed horizontal stripes [38, 43, 52, 55]. (2) The auto-localization based methods, which try to locate human parts through the learned grids [20, 51, 54]. (3) The attention based methods, which construct the part alignment through enhancing the discriminative regions and suppressing the background [23, 3, 50, 58]. Most of the above methods are coarse with much background noise in their located parts and do not consider the situation that some human parts disappear in an image due to occlusion. The first row of Figure 1 illustrates these streams.

The extra semantic based methods inject extra semantic in terms of part/pose to achieve the part alignment at pixel-level [18, 25, 32, 53]. Their success heavily counts on the accuracy of the extra pre-trained human parsing models or pose estimators. Most importantly, the identifiable personal belongings (e.g., backpacks and reticule), which are the potentially useful contextual cues for identifying a person, cannot be recognized by these pre-trained models and discarded as background. The failure cases of the extra semantic based methods are shown in the second row of Figure 1.

In this paper, we propose an extra semantic free method, Identity-guided Semantic Parsing (ISP), which can locate both human body parts and potential personal belongings at pixel-level only with the person identity labels. Specifically, we design the cascaded clustering on feature maps and regard the cluster

assignments as the pseudo-labels of human parts to supervise the part estimation. For the pixels of all images of a person, we first group them to foreground or background according to their activations on feature maps, basing on the reasonable assumption that the classification networks are more responsive to the foreground pixels than the background ones [42, 48, 47]. In this stage, the foreground parts are automatically searched by the network itself rather than manually predefined, and the self-learned scheme can capture the potentially useful semantic of both human body parts and personal belongings.

Next, we need to assign the human part labels to the foreground pixels. The difficulty of this stage lies in how to guarantee the semantic consistency across different images in terms of the appearance/pose variations, and especially the occlusion, which has not been well studied in previous extra semantic free approaches. To overcome this difficulty, we cluster the foreground pixels of all the images with the same ID, rather than those of a single image, into human parts (e.g., head, backpacks, upper-body, legs and shoes), so that the number of assigned semantic parts of a single image can adaptively vary when the instance is occluded. Consequently, our scheme is robust to the occlusion and the assigned pseudo-labels of human parts across different images are ensured to be semantically consistent. The second row of Figure 1 shows the assigned pseudo-labels.

We iteratively cluster the pixels of feature maps and employ the cluster assignments as pseudo-labels of human parts to learn the part representations. In this iterative mechanism, the generated pseudo-labels become finer and finer, resulting in the more and more accurate part estimation. The predicted probability maps of part estimation are then used to conduct the part pooling for partial representations of both human body parts and personal belongings. During matching, we only consider local features of the shared-visible parts between probe and gallery images. Besides, ISP is a generally applicable and backbone-agnostic approach, which can be readily applied in popular networks.

We summarize the contributions of this work as follows:

- In this paper, we propose the identity-guided human semantic parsing approach (ISP) for aligned person re-ID, which can locate both the human body parts and personal belongings (e.g., backpacks and reticule) at pixel-level only with the image-level supervision of person identities.
- To the best of our knowledge, ISP is the first extra semantic free method that can explicitly identify the visible parts from the occluded images. The occluded parts are excluded and only features of the shared-visible parts between probe and gallery images are considered during the feature matching.
- We set the new state-of-the-art performance on three person re-ID datasets, Market-1501 [57], DukeMTMC-reID [59] and CUHK03-NP [22, 60].

2 Related work

2.1 Semantic learning with image-level supervision

To the best of our knowledge, there is no previous work to learning human semantic parsing with image-level supervision but only weakly-supervised methods for

semantic segmentation [16, 47, 19, 12, 11, 33, 44, 62], which aim to locate objects like person, horse or dog at pixel-level with image-level supervision. However, all these methods cannot be used for the weakly-supervised human parsing task because they focus on different levels. Besides, their complex network structures and objective functions are not suitable for the end-to-end learning of person re-ID. Therefore, we draw little inspiration from these methods.

2.2 Alignment-based person re-ID

The alignment-baed methods can be roughly summarized to the four streams:

Rigid stripe based approaches. Some researchers directly partition the person image into rigid horizontal stripes to learn local features [38, 43, 52, 55]. Wang et al. [43] design a multiple granularity network, which contains horizontal stripes of different granularities. Zhang et al. [52] introduce a shortest path loss to align rigidly divided local stripes. However, the stripe-based partition is too coarse to well align the human parts and introduces lots of background noise.

Auto-localization based approaches. A few works have been proposed to automatically locate the discriminative parts by incorporating a regional selection sub-network [20, 51]. Li et al. [20] exploit the STN [17] for locating latent parts and subsequently extract aligned part features. However, the located grids of latent parts are still coarse and with much overlap. Besides, they produce a fixed number of latent parts, which cannot handle the occluded images.

Attention based approaches. Attention mechanism constructs alignment by suppressing background noise and enhancing the discriminative regions [3, 23, 31, 41, 50, 58]. However, these methods cannot explicitly locate the semantic parts and the consistency of focus area between images is not guaranteed.

Extra semantic based approaches. Many works employ extra semantic in terms of part/pose to locate body parts [6, 27, 29, 30, 32, 34, 40, 46, 49, 56] and try to achieve the pixel-level alignment. Kalayeh et al. [18] propose to employ pre-trained human parsing model to provide extra semantic. Zhang et al. [53] further adopt DensePose [1] to get densely semantic of 24 regions for a person. However, the requiring of extra semantic limits the utility and robustness of these methods. First, the off-the-shelf models can make mistakes in semantic estimation and these methods cannot recorrect the mistakes throughout the training. Second, the identifiable personal belongings like backpacks and reticule, which are crucial for person re-ID, cannot be recognized and ignored as background.

In this paper, we adopt the clustering to learn the human semantic parsing only with person identity labels, which can locate both human body parts and personal belongings at pixel-level. Clustering is a classical unsupervised learning method that groups similar features, while its capability has not been fully explored in the end-to-end training of deep neural networks. Recently, Mathilde et al. [2] adopt clustering to the end-to-end unsupervised learning of image classification. Lin et al. [26] also use clustering to solve the unsupervised person re-ID task. Different from them, we go further by grouping pixels to human parts to generate the pseudo-part-labels at pixel-level, which is more challenging due to

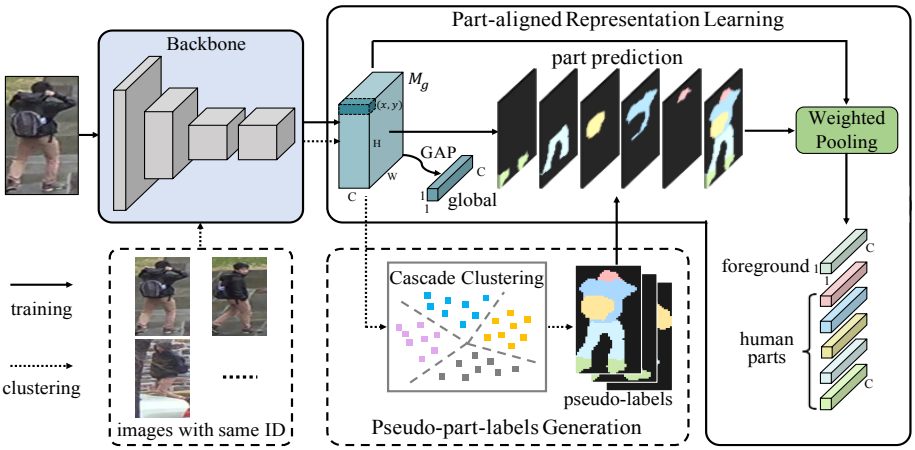

Fig. 2. The overview of ISP. The *solid line* represents the training phase and the *dotted line* represents the clustering phase. The two stages are iteratively done until the network converges. ISP is a generally applicable and backbone-agnostic approach.

various noises. Moreover, the results of clustering must guarantee the semantic consistency across images.

3 Methodology

The overview of ISP is shown in Figure 2. There are mainly two processes in our approach, i.e., pseudo-part-labels generation and part-aligned representation learning. We repeat the above two processes until the network converges.

3.1 Pixel-level part-aligned representation learning

Given n training person images $\{X_i\}_{i=1}^n$ from n_{id} distinct people and their identity labels $\{y_i\}_{i=1}^n$ (where $y_i \in \{1, ..., n_{id}\}$), we could learn the human semantic parsing to obtain the pixel-level part-aligned representations for person re-ID. For image x_i, the backbone mapping function (defined as f_θ) will output the global feature map:

$$M_g^{c\times h\times w} = f_\theta(x_i) \tag{1}$$

where θ is the parameters of backbone, and c, h, w is the channel, height and width. For clear exposition, we omit the channel dimension and denote by $M_g(x, y)$ the feature at spatial position (x, y), which is a vector of c-dim.

The main idea of our pixel-level part-aligned representations is to represent human parts with the representations of pixels belonging to that part, which is the aggregation of the pixel-wise representations weighted by a set of confidence maps. Each confidence map is used to surrogate a human part. Assuming there are $K - 1$ human parts and one background part in total, we need to estimate

K confidence maps of different semantic parts for every person image. It should be noted that we treat the personal belongings as one category of human parts. The K confidence maps is defined as $P_0, P_1, ..., P_{K-1}$, where each confidence map P_k is associated with a semantic part. We denote by $P_k(x, y)$ the confidence of pixel (x, y) belonging to semantic part k. Then the feature map of part k can be extracted from the global feature map by:

$$M_k = P_k \circ M_g \tag{2}$$

where $k \in \{0, ..., K-1\}$ and $\circ$ is the element-wise product. Adding M_k from $k = 1$ to $k = K-1$ in element-wise will get the foreground feature map M_f. Ideally, for the occluded part k in an occluded person image, $\forall_{(x,y)} P_k(x, y) = 0$ should be satisfied, which is reasonable that the network should not produce representations for the invisible parts.

3.2 Cascaded clustering for pseudo-part-labels generation

Existing studies integrate human parsing results to help capture the human body parts at pixel-level [18, 25, 32]. However, there are still many useful contextual cues like backpacks and reticule that do not fall into the scope of manually predefined human body parts. We design the cascaded clustering on feature maps M_g to generate the pseudo-labels of human parts, which includes both human body parts and personal belongings.

Specifically, in the first stage, for all M_g of the same person, we group their pixels into the foreground or background according to the activation, basing on the conception that the foreground pixels have a higher response than background ones [42, 48, 47]. In this stage, the discriminative foreground parts are automatically searched by the network and the self-learned scheme could apply both the human body parts and the potential useful personal belongings with high response. We regard the l_2-norm of $M_g(x, y)$ as the activation of pixel (x, y). For all pixels of a M_g, we normalize their activations with their maximum:

$$a(x, y) = \frac{||M_g(x, y)||_2}{\max_{(i,j)} ||M_g(i, j)||_2} \tag{3}$$

where (i, j) is the positions in the M_g and the maximum of $a(x, y)$ equals to 1.

In the second stage, we cluster all the foreground pixels assigned by the first clustering step into $K-1$ semantic parts. The number of semantic parts for a single image could be less than $K-1$ when the person is occluded because the cluster samples are foreground pixels of all M_g from the images of the same person, rather than M_g of a single image. Therefore, the clustering is robust to the occlusion and the part assignments across different images are ensured to be semantically consistent. In this stage, we focus on the similarities and differences between pixels rather than activation thus l_2-normalization is used:

$$D(x, y) = \frac{M_g(x, y)}{||M_g(x, y)||_2} \tag{4}$$

The cluster assignments are then used as the pseudo-labels of human parts, which contain the personal belongings as one foreground part, to supervise the learning of human semantic parsing. We assign label 0 to background and the body parts are assigned to label $\{1, ..., K-1\}$ according to the average position from top to down. ISP iteratively does the cascade clustering on feature maps and uses the assignments as pseudo-part-labels to learn the partial representations. In this iterative mechanism, the generated pseudo-labels become finer and finer, resulting in more and more accurate part estimation for aligned person re-ID.

Optimization. For part prediction, we use a linear layer followed by softmax activation as the classifier, which is formulated as:

$$P_k(x,y) = softmax(W_k^T M_g(x,y)) = \frac{\exp(W_k^T M_g(x,y))}{\sum_{i=0}^{K-1} \exp(W_i^T M_g(x,y))} \tag{5}$$

where $k \in \{0, ..., K-1\}$ and W is the parameters of linear layer.

We assign the probability $P_k(x,y)$ as the confidence of pixel (x,y) belonging to the semantic part k and employ cross-entropy loss to optimize the classifier:

$$\mathcal{L}_{parsing} = \sum_{x,y} -\log P_{k_i}(x,y) \tag{6}$$

where k_i is the generated pseudo-label of human parts for pixel (x,y).

3.3 Objective function

The representation for semantic part k is obtained by $F_k = GAP(M_k)$, where GAP means global average pooling. We concatenate all F_k except $k=0$ and regard the outcome as a whole representation of local parts for training. Besides, the representations for foreground and global image are directly obtained by $F_f = GAP(M_f)$, $F_g = GAP(M_g)$. In fact, the probability map product together with the GAP is the operation of weighted pooling as indicated in Figure 2.

In the training phase, we employ three groups of basic losses for the representations of local part, foreground and global image separately, which are denoted as $\mathcal{L}_p$, $\mathcal{L}_f$ and $\mathcal{L}_g$. For each basic loss group, we follow [28] to combine the triplet loss [10] and cross-entropy loss with label smoothing [39]. Therefore, the overall objective function is:

$$\mathcal{L}_{reid} = \mathcal{L}_p + \mathcal{L}_f + \mathcal{L}_g + \alpha \mathcal{L}_{parsing} \tag{7}$$

where α is the balanced weight and is set to 0.1 in our experiments.

3.4 Aligned representation matching

As illustrated in Figure 3, the final distance between query and gallery images consists of two parts. One is the distance of global and foreground features, which always exist. The other is the distance of the partial features between the shared-visible human parts. The matching strategy is inspired by [29], but [29] utilizes

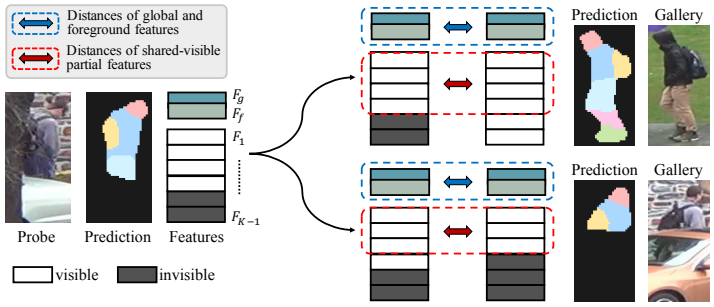

Fig. 3. The matching strategy of ISP. The distance between probe and gallery images are measured by features of the global image and foreground part, which always exist, and the features of shared-visible parts.

extra pose information and only achieves stripe-level alignment, while we do not require any extra semantic and could identify the visible parts at pixel-level. As the $\text{argmax}_i P_i(x, y)$ indicates the semantic part of pixel (x, y) belonging to, we could easily obtain the label of whether part k is visible $l_k \in \{0, 1\}$ by:

$$l_k = \begin{cases} 1, & \text{if } \exists (x,y) \in \{(x,y) | \text{argmax}_i P_i(x,y) = k\} \\ 0, & \text{else} \end{cases} \quad (i = 0, ..., K-1) \tag{8}$$

Now the distance d_k of the kth part between query and gallery images is:

$$d_k = D(F_k^q, F_k^g) \quad (k = 1, ..., K-1) \tag{9}$$

where $D()$ denotes the distance metric, which is cosine distance in this paper. F_k^q, F_k^g denote the kth partial feature of the query and gallery image, respectively. Similarly, the measure distance between global and foreground features are formulated as: $d_g = D(F_g^q, F_g^g)$, $d_f = D(F_f^q, F_f^g)$. Then, the final distance d could be obtained by:

$$d = \frac{\sum_{k=1}^{K-1} (l_k^q \cdot l_k^g) d_k + (d_g + d_f)}{\sum_{k=1}^{K-1} (l_k^q \cdot l_k^g) + 2} \tag{10}$$

If the kth parts of both the query and gallery images are visible, $l_k^q \cdot l_k^g = 1$. Else, $l_k^q \cdot l_k^g = 0$. To the best of our knowledge, ISP is the first extra semantic free method that explicitly addresses the occlusion problem.

4 Experiments

4.1 Datasets and evaluation metrics

Holistic person re-ID datasets. We select three widely used holistic person re-ID benchmarks, Market-1501 [57] which contains 32668 person images of 1501

identities, DukeMTMC-reID [59] which contains 36411 person images of 1402 identities and CUHK03-NP (New Protocol) [22, 60] which contains 14097 person images of 1467 identities for evaluation. Following common practices, we use the cumulative matching characteristics (CMC) at Rank-1, Rank-5, and the mean average precision (mAP) to evaluate the performance.

Occluded person re-ID datasets. We also evaluate the performance of ISP in the occlusion scenario. Occluded-DukeMTMC [29], which contains 15618 training images, 17661 gallery images, and 2210 occluded query images, is by far the largest and the only occluded person re-ID datasets that contains training set. It is a new split of DukeMTMC-reID [59] and the training/query/gallery set contains 9%/100%/10% occluded images, respectively. Therefore, we demonstrate the effectiveness of ISP in occluded scenario on this dataset.

4.2 Implementation details

Data preprocessing. The input images are resized to 256×128 and the global feature map M_g is 1/4 of the input size. As for data augmentation, we adopt the commonly used random cropping [45], horizontal flipping and random erasing [41, 45, 61] (with a probability of 0.5) in both the baseline and our schemes.

Optimization. The backbone network is initialized with the pre-trained parameters on ImageNet [4]. We warm up the model for 10 epochs with a linearly growing learning rate from 3.5×10^{-5} to 3.5×10^{-4}. Then, the learning rate is decreased by a factor of 0.1 at 40th and 70th epoch. We observe that 120 epochs are enough for model converging. The batch size is set to 64 and adam method is adopted to optimize the model. All our methods are implemented on PyTorch.

Clustering for reassignment. We adopt k-means as our clustering algorithm and reassign the clusters every n epochs, which is a tradeoff between the parameter updating and the pseudo-label generation. We find out that simply setting $n = 1$ is nearly optimal. We do not define any initial pseudo-labels for person images and the first clustering is directly conducted on the feature maps output by the initialized backbone. As for the time consumption, to train a model, the overall clustering time is about 6.3h/5.4h/3.1h for datasets of DukeMTMC-reID/Market1501/CUHK03-NP through multi-processes with one NVIDIA TITAN X GPU. Most importantly, the testing time is not increased at all.

4.3 Comparison with state-of-the-art methods

We compare our method with the state-of-the-art methods for holistic and occluded person re-ID in Table 1 and Table 2, respectively.

DukeMTMC-reID. ISP achieves the best performance and outperforms others by at least 0.5%/1.6% in Rank-1/mAP. On this dataset, the semantic extracted by pre-trained model is error-prone [53], which leads to significant performance degradation for extra semantic based methods. This also proves the learned semantic parts are superior to the outside ones in robustness.

Market1501. ISP achieves the best performance on mAP accuracy and the second best on Rank-1. We further find that the improvement of mAP score

Table 1. Comparison with state-of-the-art methods of the holistic re-ID problem. The $1^{st}/2^{nd}$ results are shown in red/blue, respectively. The methods in the 1st group are rigid stripe based. The methods in the 2nd group are auto-localization based. The 3rd group is attention based methods. The methods in the 4th group are extra semantic based. The last line is our method

Methods	Ref	DukeMTMC			Market1501			CUHK03-NP			
								Labeled		Detected	
		R-1	R-5	mAP	R-1	R-5	mAP	R-1	mAP	R-1	mAP
AlignedReID [52]	Arxiv18	-	-	-	91.8	97.1	79.3	-	-	-	-
PCB+RPP [38]	ECCV18	83.3	-	69.2	93.8	97.5	81.6	-	-	63.7	57.5
MGN [43]	MM18	88.7	-	78.4	95.7	-	86.9	68.0	67.4	66.8	66.0
MSCAN [20]	CVPR17	-	-	-	80.8	-	57.5	-	-	-	-
PAR [54]	ICCV17	-	-	-	81.0	92.0	63.4	-	-	-	-
DuATM [31]	CVPR18	81.8	90.2	64.6	91.4	97.1	76.6	-	-	-	-
Mancs [41]	ECCV18	84.9	-	71.8	93.1	-	82.3	69.0	63.9	65.5	60.5
IANet [13]	CVPR19	87.1	-	73.4	94.4	-	83.1	-	-	-	-
CASN+PCB [58]	CVPR19	87.7	-	73.7	94.4	-	82.8	73.7	68.0	71.5	64.4
CAMA [50]	CVPR19	85.8	-	72.9	94.7	98.1	84.5	70.1	66.5	66.6	64.2
MHN-6 [3]	ICCV19	89.1	94.6	77.2	95.1	98.1	85.0	77.2	72.4	71.7	65.4
SPReID [18]	CVPR18	84.4	-	71.0	92.5	-	81.3	-	-	-	-
PABR [35]	ECCV18	84.4	92.2	69.3	91.7	96.9	79.6	-	-	-	-
AANet [40]	CVPR19	87.7	-	74.3	93.9	-	83.4	-	-	-	-
DSA-reID [53]	CVPR19	86.2	-	74.3	95.7	-	87.6	78.9	75.2	78.2	73.1
P^2-Net [6]	ICCV19	86.5	93.1	73.1	95.2	98.2	85.6	78.3	73.6	74.9	68.9
PGFA [29]	ICCV19	82.6	-	65.5	91.2	-	76.8	-	-	-	-
ISP (ours)	ECCV20	89.6	95.5	80.0	95.3	98.6	88.6	76.5	74.1	75.2	71.4

brought by ISP is larger than that of Rank-1, which indicates ISP effectively advances the ranking positions of misaligned person images as mAP is a comprehensive index that considers all the ranking positions of the target images.

CUHK03-NP. ISP achieves the second best results. In CUHK03-NP, a great many of images contain incomplete person bodies and some human parts even disappear from all the images of a person. But ISP requires at least every semantic part appears once for a person to guarantee a high consistency on semantic. Even so, ISP still outperforms all other approaches except DSA-reID [53] which employs extra supervision by a pre-trained DensePose model [1], while our method learns the pixel-level semantic without any extra supervision.

Occluded-DukeMTMC. ISP sets a new state-of-the-art performance and outperforms others by a large margin, at least 11.2%/14.3% in Rank-1/mAP. ISP could explicitly identify the visible parts at pixel-level from the occluded images and only the shared-visible parts between query and gallery images are considered during the feature matching, which greatly improves the performance. As shown in Table 2, the aligned representation matching strategy brings considerable improvement, e.g., 3.3% of Rank-1 and 0.9% of mAP.

Table 2. Comparison with state-of-the-art methods of the occluded re-ID problem on Occluded-DukeMTMC. Methods in the 1st group are for the holistic re-ID problem. Methods in the 2nd group utilize extra pose information for occluded re-ID problem. The methods in the 3rd group do not adopt extra semantic. The last line is our method

Methods	Rank-1	Rank-5	Rank-10	mAP
HACNN [24]	34.4	51.9	59.4	26.0
Adver Occluded [15]	44.5	-	-	32.2
PCB [38]	42.6	57.1	62.9	33.7
Part Bilinear [36]	36.9	-	-	-
FD-GAN [5]	40.8	-	-	-
PGFA [29]	51.4	68.6	74.9	37.3
DSR [8]	40.8	58.2	65.2	30.4
SFR [9]	42.3	60.3	67.3	32.0
$ISP_{w/o\ arm}$	59.5	73.5	78.0	51.4
ISP (ours)	**62.8**	**78.1**	**82.9**	**52.3**

4.4 The performance of the learned human semantic parsing

As there are no manual part labels for the person re-ID datasets, we adopt the state-of-the-art parsing model SCHP [21] pre-trained on Look into Person (LIP) [25] to create the "ground-truth" of four parts.[4] Then we adopt segmentation Intersection over Union (IoU) to evaluate both the accuracy of the pseudo-part-labels of the training set and that of the semantic estimation on the testing set, which are detailed in Table 3, and the results are high enough for the IoU evaluation metric. We also find an interesting phenomenon that the accuracy of the predicted semantic estimation on testing set is mostly higher than that of the pseudo-part-labels for the training set, which indicates that, with the person re-ID supervision, our part estimator is robust to the false pseudo-labels and obtains an enhanced generalization capability.

Table 3. The human semantic parsing performance (%) of ISP ($K=6$)

IoU	Datasets	Foreground	Head	Legs	Shoes
Pseudo-labels Accuracy	DukeMTMC	65.66	68.17	61.83	58.89
	Market1501	65.45	54.74	67.02	55.25
	CUHK03-NP-Labeled	51.26	68.21	52.24	57.60
Prediction Accuracy	DukeMTMC	66.94	71.35	68.02	62.60
	Market1501	63.44	55.78	69.10	56.32
	CUHK03-NP-Labeled	53.51	59.96	50.14	59.08

[4] The parts of hat, hair, sunglass, face in LIP is aggregated as the "ground-truth" for Head; the parts of left-leg, right-leg, socks, pants are aggregated as Legs; the parts of left-shoe and right-shoe are aggregated as Shoes.

We further conduct three visualization experiments to show the effect of ISP. First, we visualize our pseudo-part-labels under different K and show the comparison with SCHP [21] in Figure 4, which validates that ISP can recognize the personal belongings (e.g. backpacks and reticule) as one human part while the pre-trained parsing model cannot. The first row of Figure 4 shows our capability of explicitly locating the visible parts in occluded images. Moreover, Figure 4 also validates the semantic consistency in ISP.

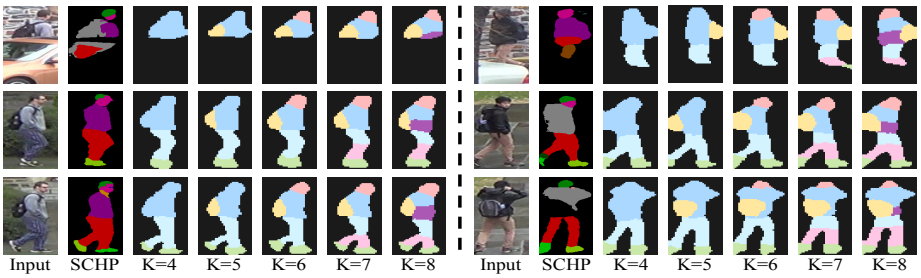

Fig. 4. The assigned pseudo-labels of human parts. From left to right: input images, semantic estimation by SCHP [21], the assigned pseudo-labels with different K.

Second, to validate the necessity of the cascaded clustering, we compare the pseudo-part-labels by cascaded clustering with Variant 1 which directly clusters the semantic in one step, and Variant 2 which removes the l_2 normalization. Figure 5 indicates that the alignment of Variant 1 is coarse and error-prone, and Variant 2 assigns an unreasonable semantic part with sub-response surround human bodies, which indicates the clustering is influenced by the activation.

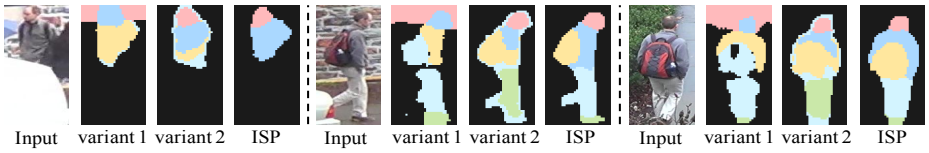

Fig. 5. The necessity of the operations of cascaded clustering ($K = 6$).

Finally, Figure 6 shows the evolution process of the pseudo-part-labels ($K = 6$), which presents a clear process of coarse-to-fine. The first clustering is directly conducted on feature maps output by the initialized network.

4.5 Ablation studies on re-ID performance

Choice of K clustering categories. Intuitively, the number of clustering centers K determines the granularity of the aligned parts. We perform the quanti-

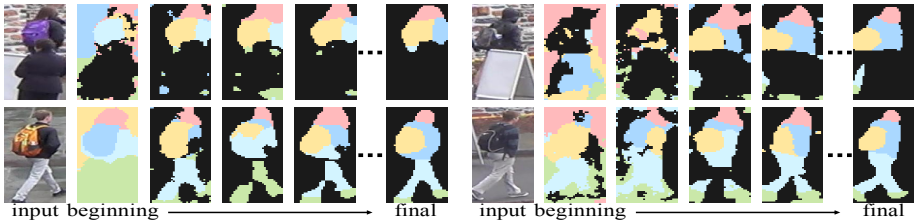

Fig. 6. The evolution process of pseudo-labels, which is a gradual process of refinement.

Table 4. The ablation studies of K. The results show ISP is robust to different K

K	DukeMTMC			Market1501			CUHK03-NP			
							Labeled		Detected	
	R-1	R-5	mAP	R-1	R-5	mAP	R-1	mAP	R-1	mAP
4	89.6	95.2	79.2	**95.3**	**98.6**	**88.6**	75.9	72.9	73.5	71.3
5	88.6	95.0	79.1	94.4	98.2	87.7	73.9	72.4	73.1	71.0
6	89.0	95.1	78.9	95.2	98.4	88.4	75.9	73.8	**75.2**	**71.4**
7	**89.6**	**95.5**	**80.0**	95.0	98.2	88.4	**76.5**	**74.1**	73.6	70.8
8	88.9	94.7	78.4	94.9	98.5	88.6	75.9	73.1	74.0	71.4

tative ablation studies to clearly find the most suitable K. As detailed in Table 4, the performance of ISP is robust to different K. Besides, we also find that $K=5$ is always the worst, which is consistent with its lowest accuracy of semantic parsing. For example, $K=5$ only obtains the pseudo-labels accuracy (IoU) of 64.25%, 53.23% and 54.83% for foreground, legs and shoes on DukeMTMC-reID.

Learned semantic vs. extra semantic. We further conduct experiments to validate the superiority of the learned semantic over extra semantic. HRNet-W32 [37] is set as our baseline model. "+extra info" means adopting the extra semantic information extracted by SCHP [21] as the human part labels while "ISP" adopts the learned semantic. The results are list in Table 5, which show ISP consistently outperforms "+extra info" method by a considerable margin. We think this is mainly because: (1) ISP can recognize the identifiable personal belongs while "+extra info" cannot. (2) "+extra info" cannot recorrect the semantic estimation errors throughout the training while ISP can recorrect its mistakes every epoch, thus ISP is less likely to miss the key clues.

Choice of backbone architecture. As ISP is a backbone-agnostic approach, we show the effectiveness of ISP with different backbones including ResNet [7], SeResNet [14] and HRNet [37]. A bilinear upsample layer is added to scale up the final feature maps of ResNet50 and SeResNet50 to the same size of HRNet-W32. As Table 6 shows, HRNet-W32 obtains the highest performance. We think it is because HRNet maintains high-resolution representations throughout the network, which could contain more semantic information.

The matching results. We compare the ranking results of baseline and ISP in Figure 7, which indicates ISP can well overcome the misalignment problem in-

Table 5. The comparison of learned semantic and extra semantic

Model	DukeMTMC			Market1501			CUHK03-NP			
							Labeled		Detected	
	R-1	R-5	mAP	R-1	R-5	mAP	R-1	mAP	R-1	mAP
baseline	87.7	94.3	77.2	94.0	97.9	85.9	71.9	68.5	67.6	64.7
+extra info	88.6	94.7	79.1	94.8	98.4	87.7	73.3	71.9	72.2	69.6
ISP	89.6	95.5	80.0	95.3	98.6	88.6	76.5	74.1	75.2	71.4

Table 6. The ablation studies of different backbone networks on DukeMTMC-reID.

backbone	#params	R-1	R-5	mAP
HRNet-W32	28.5M	89.6	95.5	80.0
ResNet50	25.6M	88.7	94.9	78.9
SeResNet50	28.1M	88.8	95.2	79.2

cluding part occlusions, inaccurate person detection, and human pose variations. Besides, we can also observe the benefit of identifying the personal belongings.

Fig. 7. The ranking results of baseline (the first row) and ISP (the second row).

5 Conclusion

In this paper, we propose the identity-guided human semantic parsing method for aligned person re-identification, which can locate both human body parts and personal belongings at pixel-level only with image-level supervision of person identities. Extensive experiments validate the superiority of our method.

Acknowledgement. This work was supported by Key-Area Research and Development Program of Guangdong Province (No.2020B010165001), National Natural Science Foundation of China (No.61772527, 61976210, 61702510), China Postdoctoral Science Foundation No.2019M660859, Open Project of Key Laboratory of Ministry of Public Security for Road Traffic Safety (No.2020ZDSYSKFKT04).

References

1. Alp Güler, R., Neverova, N., Kokkinos, I.: Densepose: Dense human pose estimation in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7297–7306 (2018)
2. Caron, M., Bojanowski, P., Joulin, A., Douze, M.: Deep clustering for unsupervised learning of visual features. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 132–149 (2018)
3. Chen, B., Deng, W., Hu, J.: Mixed high-order attention network for person re-identification. In: Proceedings of the IEEE International Conference on Computer Vision (2019)
4. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
5. Ge, Y., Li, Z., Zhao, H., Yin, G., Yi, S., Wang, X., et al.: Fd-gan: Pose-guided feature distilling gan for robust person re-identification. In: Advances in neural information processing systems. pp. 1222–1233 (2018)
6. Guo, J., Yuan, Y., Huang, L., Zhang, C., Yao, J.G., Han, K.: Beyond human parts: Dual part-aligned representations for person re-identification. In: The IEEE International Conference on Computer Vision (ICCV) (October 2019)
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
8. 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)
9. He, L., Sun, Z., Zhu, Y., Wang, Y.: Recognizing partial biometric patterns. arXiv preprint arXiv:1810.07399 (2018)
10. Hermans, A., Beyer, L., Leibe, B.: In defense of the triplet loss for person re-identification. arXiv preprint arXiv:1703.07737 (2017)
11. Hong, S., Noh, H., Han, B.: Decoupled deep neural network for semi-supervised semantic segmentation. In: Advances in neural information processing systems. pp. 1495–1503 (2015)
12. Hong, S., Yeo, D., Kwak, S., Lee, H., Han, B.: Weakly supervised semantic segmentation using web-crawled videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7322–7330 (2017)
13. Hou, R., Ma, B., Chang, H., Gu, X., Shan, S., Chen, X.: Interaction-and-aggregation network for person re-identification. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
14. 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)
15. Huang, H., Li, D., Zhang, Z., Chen, X., Huang, K.: Adversarially occluded samples for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5098–5107 (2018)
16. Huang, Z., Wang, X., Wang, J., Liu, W., Wang, J.: Weakly-supervised semantic segmentation network with deep seeded region growing. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7014–7023 (2018)
17. Jaderberg, M., Simonyan, K., Zisserman, A., et al.: Spatial transformer networks. In: Advances in neural information processing systems. pp. 2017–2025 (2015)

18. Kalayeh, M.M., Basaran, E., Gökmen, M., Kamasak, M.E., Shah, M.: Human semantic parsing for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1062–1071 (2018)
19. Lee, J., Kim, E., Lee, S., Lee, J., Yoon, S.: Ficklenet: Weakly and semi-supervised semantic image segmentation using stochastic inference. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5267–5276 (2019)
20. Li, D., Chen, X., Zhang, Z., Huang, K.: Learning deep context-aware features over body and latent parts for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 384–393 (2017)
21. Li, P., Xu, Y., Wei, Y., Yang, Y.: Self-correction for human parsing. arXiv preprint arXiv:1910.09777 (2019)
22. Li, W., Zhao, R., Xiao, T., Wang, X.: Deepreid: Deep filter pairing neural network for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 152–159 (2014)
23. Li, W., Zhu, X., Gong, S.: Harmonious attention network for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2285–2294 (2018)
24. Li, W., Zhu, X., Gong, S.: Harmonious attention network for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2285–2294 (2018)
25. Liang, X., Gong, K., Shen, X., Lin, L.: Look into person: Joint body parsing & pose estimation network and a new benchmark. IEEE transactions on pattern analysis and machine intelligence **41**(4), 871–885 (2018)
26. Lin, Y., Dong, X., Zheng, L., Yan, Y., Yang, Y.: A bottom-up clustering approach to unsupervised person re-identification. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 8738–8745 (2019)
27. Liu, J., Ni, B., Yan, Y., Zhou, P., Cheng, S., Hu, J.: Pose transferrable person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4099–4108 (2018)
28. 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 Conference on Computer Vision and Pattern Recognition Workshops. pp. 0–0 (2019)
29. 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)
30. Saquib Sarfraz, M., Schumann, A., Eberle, A., Stiefelhagen, R.: A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 420–429 (2018)
31. Si, J., Zhang, H., Li, C.G., Kuen, J., Kong, X., Kot, A.C., Wang, G.: Dual attention matching network for context-aware feature sequence based person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5363–5372 (2018)
32. Song, C., Huang, Y., Ouyang, W., Wang, L.: Mask-guided contrastive attention model for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1179–1188 (2018)
33. Souly, N., Spampinato, C., Shah, M.: Semi and weakly supervised semantic segmentation using generative adversarial network. arXiv preprint arXiv:1703.09695 (2017)

34. Su, C., Li, J., Zhang, S., Xing, J., Gao, W., Tian, Q.: Pose-driven deep convolutional model for person re-identification. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3960–3969 (2017)
35. 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)
36. 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)
37. Sun, K., Xiao, B., Liu, D., Wang, J.: Deep high-resolution representation learning for human pose estimation. In: CVPR (2019)
38. 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)
39. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2818–2826 (2016)
40. Tay, C.P., Roy, S., Yap, K.H.: Aanet: Attribute attention network for person re-identifications. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
41. 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)
42. Wang, F., Jiang, M., Qian, C., Yang, S., Li, C., Zhang, H., Wang, X., Tang, X.: Residual attention network for image classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3156–3164 (2017)
43. Wang, G., Yuan, Y., Chen, X., Li, J., Zhou, X.: Learning discriminative features with multiple granularities for person re-identification. In: 2018 ACM Multimedia Conference on Multimedia Conference. pp. 274–282. ACM (2018)
44. Wang, X., You, S., Li, X., Ma, H.: Weakly-supervised semantic segmentation by iteratively mining common object features. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1354–1362 (2018)
45. Wang, Y., Wang, L., You, Y., Zou, X., Chen, V., Li, S., Huang, G., Hariharan, B., Weinberger, K.Q.: Resource aware person re-identification across multiple resolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8042–8051 (2018)
46. Wei, L., Zhang, S., Yao, H., Gao, W., Tian, Q.: Glad: Global-local-alignment descriptor for pedestrian retrieval. In: Proceedings of the 25th ACM international conference on Multimedia. pp. 420–428. ACM (2017)
47. Wei, Y., Feng, J., Liang, X., Cheng, M.M., Zhao, Y., Yan, S.: Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1568–1576 (2017)
48. Woo, S., Park, J., Lee, J.Y., So Kweon, I.: Cbam: Convolutional block attention module. In: The European Conference on Computer Vision (ECCV) (September 2018)
49. Xu, J., Zhao, R., Zhu, F., Wang, H., Ouyang, W.: Attention-aware compositional network for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2119–2128 (2018)

50. Yang, W., Huang, H., Zhang, Z., Chen, X., Huang, K., Zhang, S.: Towards rich feature discovery with class activation maps augmentation for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1389–1398 (2019)
51. Yao, H., Zhang, S., Hong, R., Zhang, Y., Xu, C., Tian, Q.: Deep representation learning with part loss for person re-identification. IEEE Transactions on Image Processing **28**(6), 2860–2871 (2019)
52. Zhang, X., Luo, H., Fan, X., Xiang, W., Sun, Y., Xiao, Q., Jiang, W., Zhang, C., Sun, J.: Alignedreid: Surpassing human-level performance in person re-identification. arXiv preprint arXiv:1711.08184v2 (2018)
53. Zhang, Z., Lan, C., Zeng, W., Chen, Z.: Densely semantically aligned person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 667–676 (2019)
54. Zhao, L., Li, X., Zhuang, Y., Wang, J.: Deeply-learned part-aligned representations for person re-identification. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3219–3228 (2017)
55. Zheng, F., Deng, C., Sun, X., Jiang, X., Guo, X., Yu, Z., Huang, F., Ji, R.: Pyramidal person re-identification via multi-loss dynamic training. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
56. Zheng, L., Huang, Y., Lu, H., Yang, Y.: Pose invariant embedding for deep person re-identification. IEEE Transactions on Image Processing (2019)
57. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: A benchmark. In: Computer Vision, IEEE International Conference on Computer Vision. pp. 1116–1124 (2015)
58. Zheng, M., Karanam, S., Wu, Z., Radke, R.J.: Re-identification with consistent attentive siamese networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5735–5744 (2019)
59. Zheng, Z., Zheng, L., Yang, Y.: Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3754–3762 (2017)
60. Zhong, Z., Zheng, L., Cao, D., Li, S.: Re-ranking person re-identification with k-reciprocal encoding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1318–1327 (2017)
61. Zhong, Z., Zheng, L., Kang, G., Li, S., Yang, Y.: Random erasing data augmentation. arXiv preprint arXiv:1708.04896 (2017)
62. Zhou, Y., Zhu, Y., Ye, Q., Qiu, Q., Jiao, J.: Weakly supervised instance segmentation using class peak response. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3791–3800 (2018)