Body Part-Based Representation Learning for Occluded Person Re-Identification

Vladimir Somers
EPFL & UCLouvain &
Sportradar
vladimir.somers@epfl.ch

Christophe De Vleeschouwer
UCLouvain, Belgium
christophe.devleeschouwer
@uclouvain.be

Alexandre Alahi
EPFL, Switzerland
alexandre.alahi@epfl.ch

Abstract

Occluded person re-identification (ReID) is a person retrieval task which aims at matching occluded person images with holistic ones. For addressing occluded ReID, part-based methods have been shown beneficial as they offer fine-grained information and are well suited to represent partially visible human bodies. However, training a part-based model is a challenging task for two reasons. Firstly, individual body part appearance is not as discriminative as global appearance (two distinct IDs might have the same local appearance), this means standard ReID training objectives using identity labels are not adapted to local feature learning. Secondly, ReID datasets are not provided with human topographical annotations. In this work, we propose BPBreID, a body part-based ReID model for solving the above issues. We first design two modules for predicting body part attention maps and producing body part-based features of the ReID target. We then propose GiLt, a novel training scheme for learning part-based representations that is robust to occlusions and non-discriminative local appearance. Extensive experiments on popular holistic and occluded datasets show the effectiveness of our proposed method, which outperforms state-of-the-art methods by 0.7% *mAP and* 5.6% *rank-1 accuracy on the challenging Occluded-Duke dataset. Our code is available at* `https://github.com/VlSomers/bpbreid`.

1. Introduction

Person re-identification [34, 17], or ReID, is a person retrieval task which aims at matching an image of a person-of-interest, called the query, with other person images from a large database, called the gallery. ReID has important applications in smart cities for video-surveillance [42, 43] or sport understanding [26, 4]. Person re-identification is generally formulated as a representation learning task and is very challenging, because person images generally suffer from background clutter, inaccurate bounding boxes, pose variations, luminosity changes, poor image quality, and occlusions [17] from street objects or other people.

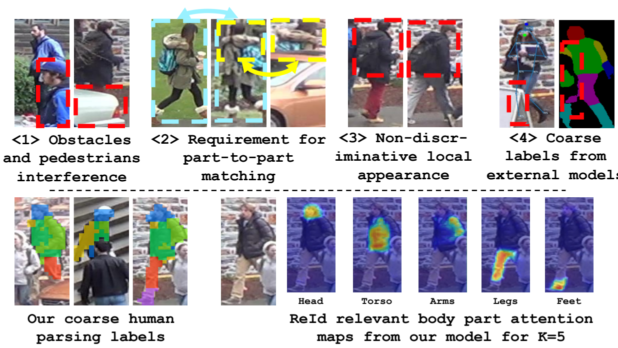

Figure 1. Overview of key concepts in our work. First row illustrates the four challenges of occluded and part-based ReID that our proposed method is trying to address. Second row illustrates our pre-generated human parsing labels and the ReID-relevant soft attention maps produced by our model BPBreID.

For solving the ReID task, most methods adopt a global approach [14, 8], learning a global representation of the target person as a single feature vector. However, these methods are unable to address the challenges caused by occlusions for two reasons, both depicted in Figure 1:

⟨1⟩ *Obstacle and pedestrians interference*: the globally learned representation might include misleading appearance information from occluding objects and pedestrians.

⟨2⟩ *Requirement for part-to-part matching*: When comparing two occluded samples, it is only relevant to compare body parts that are visible in both images. Global method cannot achieve such part-to-part matching, because the same global feature is used for every comparison.

To deal with the above issues, part-based approaches [23, 46, 37], have shown promising results. These part-based methods address the ReID task by producing multiple local feature vectors, i.e., one for each part of the input sample.

However, learning such part-based representations involves dealing with two crucial challenges:

⟨3⟩ *Non-discriminative local appearance*: Standard ReID losses, such as the id or triplet losses, work with the assumption that different identities have different appearance, and consequently that their corresponding global feature vectors are different. However, this assumption is broken when working with part-based feature vectors, because two persons with different identities might have very similar appearance on some of their body parts, as depicted in Figure 1. Because local appearance is not necessarily discriminative, standard ReID losses used for learning global representations do not scale well to local representation learning. The specificity associated to learning local features and its impact on the choice of the training loss has been overlooked in previous part-based ReID works and we are the first to point it out. To address these issues, we propose *GiLt*, a novel training loss for part-based methods. GiLt is designed to be robust to occlusions and non-discriminative local appearance, and is meant learn a set of local features that are each representative of their corresponding local parts, while being discriminative when considered jointly.

⟨4⟩ *Absence of human topology annotation*: Part-based method generally rely on spatial attention maps to perform local pooling within a global feature map and build body part features of the ReID target. However, no ReID dataset is provided with annotations regarding the local region to pool, and generating such annotation with external pose information or part segmentation tools yields inaccurate results due to the domain variation and poor image quality. Moreover, body part-based feature pooling fundamentally differ from pixel-accurate human parsing. Indeed, the spatial attention maps have to localize the body part in the image, but also to identify the feature vectors that best represent discriminant characteristics of the body part appearance. Therefore, an ideal attention map is not necessarily an accurate segmentation shape. Previous ReID works exploiting human parsing to build part-based features have either (i) used directly the output of a pose estimation model as local attention masks, without adapting it to handle the ReID task [3, 5, 15], or (ii) learned local features with part discovery, without human topology prior [13, 46, 39]. In this work, we propose a body part attention module trained with a novel dual supervision, using both identity and coarse human parsing labels. This module demonstrates how external human semantic information can be effectively leveraged to produce ReID-relevant body part attention maps.

Finally, we combine this body part attention module and GiLt (Global-identity Local-triplet) loss to build our Body Bart-Based ReID model called BPBreID, which effectively addresses all four challenges introduced before. We summarize the main contributions of our work as follows:

1. For the ReID task, we are the first to propose a soft attention trained from a dual supervision, to leverage both identity and prior human topology information. Our work demonstrates that this approach outperforms all previous part-based methods.
2. We propose a novel *GiLt* strategy for training part-based method. GiLt is robust to occlusions and non-discriminative local appearance and is straightforward to integrate with any part-based framework.
3. BPBreID outperforms state-of-the-art methods by 0.7% mAP and 5.6% rank-1 on the Occluded-Duke dataset. Our BPBreID codebase has been released to encourage further research on part-based methods.

2. Related Work

Part-based feature alignment in ReID: To solve the spatial misalignment issue, several works [23, 30, 15, 20, 38, 41, 36, 5, 15, 27, 3] adopt fixed attention mechanisms, using pre-determined pixel partitions of the input image and applying part pooling for generating local feature representations. These methods achieve poor feature selection and alignment, because the resulting attention maps are not meant for pooling ReID-relevant body part features. To solve those issues, other works [21, 10, 46, 28] use attention mechanisms trained in an end-to-end fashion for generating attention maps that are specialized towards solving the ReID task. Some of these approaches [21, 10, 28] include a pose estimation backbone as a parallel branch, which is jointly trained with the appearance backbone branch in an end-to-end fashion on the ReID dataset. However, the parallel branch induces a significant computational overhead and these methods do not address the occluded ReID problem explicitly. Other part-based methods learn local features via part discovery in a self-supervised way, without human topology prior [46, 13, 39]. Such approach might introduce alignment errors, missed parts and background clutter. Different from previous works, our part-based features are built by an attention branch which (i) is trained explicitly to pool local features that are relevant for the ReID task, and (ii) leverages external human parsing labels to bias the spatial attention in focusing on prior body regions.

Local feature learning in ReID: Identity loss and batch-hard triplet loss [8] are two popular objectives for training ReID models that are also applied to part-based methods [18, 15, 27, 5, 46, 9, 25, 3] for learning local representations. Most of these methods [23, 33, 15, 18] solely apply an identity loss on each part-based features. As a consequence, they are more sensitive to non-discriminative body parts and miss out the benefits [8, 14] of the triplet loss as a complementary deep metric learning objective for ReID. To deal with incomplete information of part features, some works [46, 25] propose to apply triplet and identity losses on a combined embedding, resulting from concatenation or summation of local features. A specialized Improved Hard Triplet Loss (IHTL) is proposed in [25] for training part-based feature, but this objective cannot cope well with occluded or similar samples. Finally, [9] applies both triplet and identity losses on combined part features, but does not use holistic features during training, which renders their training scheme less robust to inaccurate body part predictions and heavy occlusions. Our proposed GiLt training procedure aim at solving above issues and addressing the lack of consensus regarding the choice of losses to adopt for training part-based methods. Finally, it is worth noting that other works [13, 5] take an opposite approach to deal with standard ReID losses being unsuitable for non-discriminative body parts appearance. They solve this by constraining each part-based feature to be discriminative on its own, by either having each of them attending simultaneously to multiple body regions [13] or by adding high-order information in each local feature via message-passing [5].

3. Methodology

The overall architecture of our model BPBreID is depicted in Figure 2. It comprises two modules: the body part attention module described in Section 3.1 and the global-local representation learning module described in Section 3.2. The overall training procedure of BPBreID is described in Section 3.3 and the procedure used at inference for computing query to gallery distance is described in Section 3.4.

3.1. Body Part Attention Module

The body part attention module takes as an input the feature map extracted by the backbone and outputs a set of attention maps highlighting the body parts of the ReID target. This module consists of a *pixel-wise part classifier* trained with a *body part attention loss* using our coarse *human parsing labels*. We detail these three components below. Because our model is trained end-to-end, the body part attention module also receives a training signal from the ReID loss, which uses the identity labels, as described in Section 3.3. This attention branch is therefore trained from a **dual supervision**, with both a body part prediction objective and a ReID objective. As a result of the dual supervision, this module generates attention maps that are more relevant to the ReID task than the attention maps we would obtain using the fixed output of a pre-trained human parsing model. This module is depicted in the top left part of Figure 2.

3.1.1 Pixel-wise Part Classifier

The body part attention module takes in input the appearance map G, which is a tensor $R^{H\times W\times C}$ produced by a feature extractor. For each pixel (w,h) in the appearance map G, a pixel-wise part classifier predicts if it belongs to the background or to one of the K body parts, which means there are $K+1$ target classes, with the class at index 0 being the background. A 1x1 convolution layer with parameters $P \in R^{(K+1)\times C}$ followed by a $softmax$ is applied on G to obtain the classification scores $M \in R^{H\times W\times(K+1)}$:

$$M = \text{softmax}(GP^T)\ . \tag{1}$$

These $K+1$ probability maps M_k indicates therefore which pixels belong to which body parts (or to the background).

3.1.2 Human Parsing Labels

Human parsing labels $Y \in R^{H\times W}$, required for training our part attention module, are generated with the PifPaf [12] pose estimation model, following a process detailed in the supplementary materials. $Y(h,w)$ is set to $\{1, ..., \text{K}\}$ if spatial location (h,w) belong to one of the K body parts or 0 for background. Human semantic regions are defined manually for a given value of K. For instance, with $K=8$, we define the following semantic regions: $\{$*head*, *left/right arm*, *torso*, *left/right leg* and *left/right feet*$\}$. These coarse human semantic parsing labels are illustrated in Figure 1 for $K=5$.

3.1.3 Body Part Attention Loss

The pixel-wise part classifier is supervised with a body part attention loss L_{pa}, which is in practice a cross-entropy loss with label smoothing [24, 1], as formulated here:

$$\begin{gathered} L_{pa} = -\sum_{k=0}^{K}\sum_{h=0}^{H-1}\sum_{w=0}^{W-1} q_k \cdot log(M_k(h,w))\ , \\ \text{with } q_k = \begin{cases} 1-\frac{N-1}{N}\varepsilon & \text{if } Y(h,w)=k \\ \frac{\varepsilon}{N} & \text{otherwise}\ , \end{cases} \end{gathered} \tag{2}$$

where the human parsing labels map Y is described in Section 3.1.2, N is the batch size, ε is the label smoothing regularization rate and $M_k(h,w)$ is the prediction probability for part k at spatial location (w,h), as described in Eq. (1).

3.2. Global-local Representation Learning Module

The global-local representation learning module takes as input the body part attention maps generated by the previous module, and outputs holistic and body part-based features of the ReID target, together with a visibility score for each part. It can be visualized in the top right part of Figure 2. Part-based representations, combined with their visibility scores, is our solution for achieving part-to-part matching, and solving challenges $\langle 1\rangle$ and $\langle 2\rangle$ from Section 1.

3.2.1 Holistic and Body Part-based Features

As described in Section 3.1.1, the body part attention module produces K spatial heatmaps highlighting the corresponding K predicted body parts of the input image. We first combine the K body part maps $\{M_1, ..., M_K\}$ in

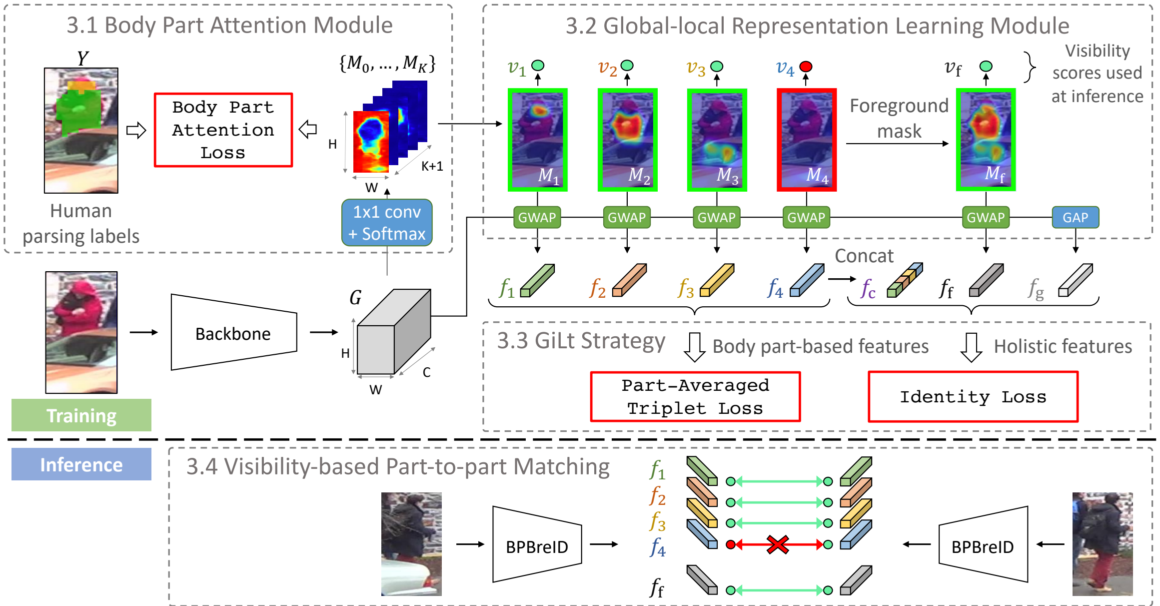

Figure 2. Structure of BPBreID with detailed architecture and training procedure in the top part, and inference procedure in bottom part. The model consists of a *body part attention module* for body part attention maps and a *global-local representation learning module* for producing holistic features $\{f_g, f_f, f_c\}$ and body part-based features $\{f_1, ..., f_K\}$ together with their visibility scores $\{v_f, v_1, ..., v_K\}$. For holistic features, "g" stands for "global", "f" for "foreground" and "c" for "concatenated". *GWAP* stands for global weighted average pooling. The network is trained in an end-to-end fashion using a *body part attention loss* for supervising part prediction, a standard *identity loss* on holistic features and a *part-averaged triplet loss* on body part-based features. Query to gallery distance is computed at inference using a *part-to-part matching strategy* for comparing only mutually visible body parts. Green/red color depict visible/invisible body parts. Each component of the architecture is framed with a grey rectangle, with its name and a number referencing the section describing it. For conciseness, BPBreID is represented here with $K = 4$: {head, torso, legs, feet}.

a single foreground heatmap $M_f \in R^{H\times W}$: $M_f(h,w) = max\big(M_1(h,w), ..., M_K(h,w)\big)$. These heatmaps are then used to perform $K+1$ *global weighted average pooling* (denoted *GWAP* in Figure 2) of the appearance feature map G, to obtain the foreground embedding f_f and the K body part-based embeddings $\{f_1, ..., f_K\}$:

$$f_i = \frac{\sum_{h=0}^{H-1}\sum_{w=0}^{W-1} G(h,w)M_i(h,w)}{\sum_{h=0}^{H-1}\sum_{w=0}^{W-1} M_i(h,w)}, \forall\, i \in \{\text{f}, 1, \ldots, K\}\,. \quad (3)$$

The initial global appearance feature map G is also globally average pooled (GAP) to obtain the global embedding f_g: $f_g = GAP(G)$. A last embedding $f_c \in R^{(C\cdot K)}$ is also produced by concatenating the K body part-based features along the channel dimension: $f_c = concat(f_1, ..., f_K)$. Our global-local representation learning module produces therefore three holistic embeddings $\{f_g, f_f, f_c\}$ and K body part-based embeddings $\{f_1, ..., f_K\}$.

3.2.2 Body Part Visibility Estimation

To detect occluded body parts, we compute a binary visibility score v_i for each embedding, with 0/1 corresponding to invisible/visible parts respectively. In our BPBreID model, visibility scores are only used at inference. For all holistic embeddings, visibility scores are set to one, i.e., $v_g = v_f = v_c = 1$. For body part-based features, visibility score v_i with $i \in \{1, ..., K\}$ is set to 1 if at least one pixel in M_i has a value above threshold λ_v, which is empirically set to 0.4, as formulated below:

$$v_i = \begin{cases} 1 & \text{if } \max\limits_{h,w}(M_i(h,w)) > \lambda_v \\ 0 & \text{otherwise}\,. \end{cases} \quad (4)$$

3.3. Overall Training Procedure

The overall objective function used to optimize the network during training stage is formulated as follows:

$$L = \lambda_{pa} L_{pa} + L_{GiLt}\ , \quad (5)$$

where L_{pa} is the body part attention loss supervised with human parsing labels (introduced in Section 3.1.3) and L_{GiLt} is our *GiLt* loss, supervised with identity labels. Parameter λ_{pa} is used to control the overall part attention loss contribution and is empirically set to 0.35.

3.3.1 GiLt Loss

To supervise model training with the identity labels, our GiLt loss relies on two losses: the popular identity classification loss and a custom part-averaged triplet loss, which is a variant of the batch hard triplet loss [8]. However, we must carefully choose which loss to apply on each of the $K+3$ embeddings produced by our model.

First, unlike other popular part-based method [23, 15, 5, 3, 32, 13, 30, 47], we do not apply the identity loss on part-based features because of occlusions and non-discriminative local appearance, as introduced in Section 1. Indeed, a part-based feature is not always discriminative enough to identify a person, which renders an identity prediction objective impossible to fulfil. Consequently, adding an identity loss on such local representation would be destructive to performance. However, similar to most state-of-the-art ReID methods, we still benefit from the identity loss supervision by applying it on the holistic features.

Secondly, we apply the triplet loss constraint on part-based features, via our custom *part-averaged triplet loss* detailed in Section 3.3.2. At inference stage, distance between samples will be computed using these part-based features, and it therefore make sense to optimize their relative distances directly with a triplet loss constraint. However, we argue the triplet constraint should not be enforced on holistic embeddings because of occlusions. Indeed, two holistic embeddings of the same identity will have intrinsically different representations if at least one of the two is partially occluded, because each embedding will represent a different subset of the whole target body. Therefore, pulling those two holistic features close together in the feature space with a triplet loss would be destructive to performance.

In summary, we claim the best training strategy for part-based methods is to apply (i) the identity loss constraint on holistic features only and (ii) the triplet loss constraint on part-based features only, via a our custom part-averaged triplet loss. We call this strategy *Global-identity Local-triplet* or simply *GiLt*, and formulate it in our GiLt loss:

$$L_{GiLt} = L_{id} + L_{tri} = \sum_{i \in \{\mathrm{g},\mathrm{f},\mathrm{c}\}} L_{CE}(f_i) + L_{tri}^{parts}(f_1, \ldots, f_K), \quad (6)$$

where L_{CE} is the cross-entropy loss with label smoothing [24] and BNNeck trick [14], and L_{tri}^{parts} is our part-averaged triplet loss detailed further below. L_{id} optimizes the network to predict the input sample identity from each holistic embedding $\{f_\mathrm{g}, f_\mathrm{f}, f_\mathrm{c}\}$.

We provide extensive ablation studies in Section 4.4 for validating our claim. These experiments also demonstrate the superiority of our GiLt strategy for training part-based methods compared to other combination of triplet and identity losses. To our knowledge, we are the first to suggest such combination of triplet and identity losses for training part-based methods. We are also the first to conduct extensive experiments to demonstrate the impact of both losses on training performance when enforced on holistic and part-based embeddings. GiLt is illustrated in Figure 2.

3.3.2 Part-Averaged Triplet Loss

Our part-averaged triplet loss differ from the standard batch hard triplet loss [8] w.r.t. the strategy used to compute the distance between two samples. Indeed, it relies on the average of pairwise parts distances between two samples i and j. This part-averaged distance is computed using all body part-based features $\{f_1, ..., f_K\}$ jointly:

$$d_{parts}^{ij} = \frac{\sum_{k=1}^{K} dist_{eucl}(f_k^i, f_k^j)}{K}, \quad (7)$$

where $dist_{eucl}$ refers to the euclidean distance. Similar to [8], the part-averaged triplet loss is then computed using the hardest positive and hardest negative part-averaged distances d_{parts}^{ap} and d_{parts}^{an} respectively:

$$L_{tri}^{parts}(f_0^a, \ldots, f_K^a) = [d_{parts}^{ap} - d_{parts}^{an} + \alpha]_+, \quad (8)$$

where the distances from anchor sample to the hardest positive and negative samples are denoted by d^{ap} and d^{an} respectively, and α is the triplet loss margin. Therefore, our part-averaged triplet loss globally optimize an average of local distances between corresponding parts, and not a distinct triplet for each part, as adopted in [5, 13] and shown to be inferior in Table 2, under "*BPBreID w/o part-averaged triplet loss*". This critical design choice gives each training step the opportunity to focus on the parts with most robust and discriminant features, which in turns mitigates the impact of occluded and non-discriminative local features.

3.4. Visibility-based Part-to-Part Matching

Given a query sample q and a gallery sample g, pairwise distance is computed at inference by a visibility-based part-to-part matching strategy using the foreground embedding and the body part-based embeddings:

$$dist_{total}^{qg} = \frac{\sum_{i \in \{\mathrm{f},1,\ldots,K\}} \left(v_i^q \cdot v_i^g \cdot dist_{eucl}(f_i^q, f_i^g) \right)}{\sum_{i \in \{\mathrm{f},1,\ldots,K\}} \left(v_i^q \cdot v_i^g \right)}. \quad (9)$$

Visibility scores $v_i^{q|g}$ are used to ensure that only mutually visible body parts are compared. If there's no mutually visible part between the two samples, their distance is set to infinity. The strategy is illustrated in the bottom part of Figure 2. Global and concatenated embeddings are not used at inference because they may convey information from occluding objects and pedestrians.

4. Experiments

4.1. Datasets and Evaluation Metrics

We evaluate our model on the holistic datasets Market-1501 [42] and DukeMTMC-reID [43], and the occluded datasets Occluded-Duke [15], Occluded-ReID[1] [48] and P-DukeMTMC [48]. We report two standards ReID metrics:

[1] Occluded-ReID has no train set, so we use Market-1501 for training.

the cumulative matching characteristics (CMC) at Rank-1 and the mean average precision (mAP). Performances are evaluated without re-ranking [44] in a single query setting.

4.2. Implementation Details

Model architecture A *ResNet-50* (RN) [6] is employed as the main backbone feature extractor. The final fully connected layer and global average pooling layer are removed and the stride of the last convolutional layer is set to 1 instead of 2. For a fair comparison with methods using heavier architecture or training, we also employ the *ResNet-50-ibn* (*RI*) [16] as in [7, 35, 31], and the *HRNet-W32* (HR) [22] backbone as in [46]. HRNet feature maps with higher resolution are particularly beneficial to BPBreID for building fine-grained attention maps. All backbones are pre-trained on ImageNet [19]. The number of body parts K is set to 5 for holistic datasets and 8 for occluded datasets. An ablation study on K is provided in the supplementary materials.

Training procedure The training procedure is mainly adopted from BoT [14]. All images are resized to 256×128 for *ResNet-50* (*RN*) and 384×128 with *HRNet-W32* (*HR*) and *ResNet-50-ibn* (*RI*). Images are first augmented with random cropping and 10 pixels padding, and then with random erasing [45] at 0.5 probability. A training batch consists of 64 samples from 16 identities with 4 images each. The model is trained in an end-to-end fashion for 120 epochs with the Adam optimizer on one NVIDIA Quadro RTX8000 GPU. The learning rate is increased linearly from 3.5×10^{-5} to 3.5×10^{-4} after 10 epochs and is decayed to 3.5×10^{-5} and 3.5×10^{-6} at 40^{th} epoch and 70^{th} epoch respectively. The label smoothing regularization rate ε is set to 0.1 and triplet loss margin α is set to 0.3.

Table 1. Comparison of BPBreID with SOTA methods. Symbols † / ‡ denote *global* / *part-based* methods respectively. First, second and third best performance are indicated with [1,2,3] respectively.

Methods	Holistic datasets				Occluded datasets					
	Market -1501		DukeMT -MC-ReID		Occluded -Duke		Occluded -reID		P-Duke -MTMC	
	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP
BoT-based training schemes with single ResNet-50 backbone										
BoT [14] †	94.5	85.9	86.4	76.4	51.4	44.7	58.4	52.3	87.0	74.9
SGAM [33] ‡	91.4	67.3	83.5	67.3	55.1	35.3	-	-	-	-
PGFA [15] ‡	91.2	76.8	82.6	65.5	51.4	37.3	-	-	44.2	23.1
MHSA [25] ‡	94.6	84.0	87.3	73.1	59.7	44.8	-	-	70.7	41.1
VGTri [32] ‡	-	-	-	-	62.2	46.3	**81.0**	**71.0**	-	-
OAMN [2] †	93.2	79.8	86.3	72.6	62.6	46.1	-	-	-	-
HG [11] †	**95.6**	86.1	87.1	77.5	61.4	50.5	-	-	-	-
$\text{BPBreID}_{\text{RN}}$ ‡	95.1	**87.0**	**89.6**	**78.3**	**66.7**	**54.1**	76.9	68.6	**91.0**	**77.8**
Arbitrary backbones/training schemes or heavier architectures										
PVPM [3] ‡	-	-	-	-	-	-	66.8	59.5	85.1	69.9
HOReID [5] ‡	94.2	84.9	86.9	75.6	55.1	43.8	80.3	70.2	-	-
ISP [46] ‡	95.3	88.6	89.6	80.0	62.8	52.3	-	-	-	-
PAT [13] ‡	95.4	88.0	88.8	78.2	64.5	53.6	81.6	72.1	-	-
PGFL [40] ‡	95.3	87.2	89.6	79.5	63.0	54.1	80.7	70.3	81.1	64.2
HPNet [9] ‡	-	-	-	-	-	-	**87.3**[1]	77.4[3]	-	-
SSGR [31] †	**96.1**[1]	89.3	91.1	81.3	69.0	57.2	78.5	72.9	-	-
FED [29] †	95.0	86.3	89.4	78.0	68.1	56.4	86.3[2]	79.3[2]	-	-
LDS [35] †	95.8[2]	**90.3**[1]	91.5[3]	82.5[3]	64.3	55.7	-	-	91.9[2]	82.9[2]
PFD [28] ‡	95.5	89.7[2]	91.2	83.2[2]	69.5[3]	61.8[2]	81.5	**83.0**[1]	-	-
$\text{BPBreID}_{\text{RI}}$ ‡	95.7	88.4	91.7[2]	81.3	71.3[2]	57.5[3]	77.0	70.9	91.3[3]	79.2[3]
$\text{BPBreID}_{\text{HR}}$ ‡	95.7[3]	89.4[3]	**92.4**[1]	**84.2**[1]	**75.1**[1]	**62.5**[1]	82.9[3]	75.2[4]	**93.0**[1]	**83.2**[1]

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

We compare our model in Table 1 with other ReID works and it ranks first overall. Methods in the first part of the table use a ResNet-50 backbone with a training procedure similar to ours and BoT [14]. Methods in the second part either use arbitrary training procedures with bigger image size [9, 3, 40, 35], more advanced backbones [28, 29, 46, 31], or heavier architecture with additional backbones or branches [28, 3, 5, 35, 40, 13].

Occluded-Duke and P-DukeMTMC: Our model outperforms all previous part-based methods (‡) on these two occluded datasets. For methods using directly the output of a pose estimation model as local attention masks [5, 15, 3, 28], the lack of end-to-end training leads to sub-optimal attention maps in terms of ReID-relevant feature pooling. For methods producing local features via part discovery [46, 13], not using prior human topology information renders their model more vulnerable to alignment errors, missed parts and background clutter. Our work demonstrates that an end-to-end training of a spatial attention branch with both identity and human parsing labels is superior to previous architecture to perform ReID-relevant part-based pooling. Recently, global methods (†) designed specifically for the occluded ReID task [11, 29, 35, 31] have shown promising performance compared to previous part-based methods. However, BPBreID outperforms all of them as well, demonstrating the advantage of part-based methods to solve the occluded task, since global methods cannot achieve part-to-part matching.

Market-1501 and DukeMTMC-ReID: Our method outperforms all part-based methods (‡) on both Market-1501 and DukeMTMC-ReID, except for PFD [28] on Market-1501, which uses a much heavier architecture, with a ViT backbone and a HRNet-W48 parallel branch for pose estimation. Regarding global methods (†), BPBreID outperform all of them on DukeMTMC-ReID, and achieves competitive performance on Market-1501, although the performance difference between most SOTA methods remains insignificant on it. This demonstrates that part-based methods are a competitive choice for holistic person ReID.

Occluded-ReID: This occluded dataset requires strong domain adaption capacity, since it does not provide a training set, and the Market-1501 dataset that is generally used for pre-training does not contain occluded samples. All

Table 2. Ablation study for the main components of BPBreID on Occluded-Duke. For the experiment "*w/o part-avgd triplet loss*", we replace our part-averaged triplet loss introduced in Eq. (8) by a distinct triplet loss for each part.

Methods	R-1	mAP
BoT [14] baseline	51.4	44.7
BPBreID	**66.7**	**54.1**
- w/o learnable attention	51.6	39.2
- w/o visibility scores	52.6	45.3
- w/o part-avgd triplet loss	64.8	51.7

Table 3. Performance comparison for body part and holistic embeddings.

Methods	R-1	mAP
f_g	60.2	47.5
f_f	64.1	49.7
f_c	64.4	50.3
f_1 (head)	47.1	24.7
f_2 (torso)	52.0	31.7
f_3 (left arm)	55.7	34.4
f_4 (right arm)	56.7	34.1
f_5 (legs)	22.3	13.3
f_6 (feet)	16.1	9.0
$\{f_1, ..., f_6\}$	65.9	52.2
$\{f_f, f_1, ..., f_6\}$	**66.1**	**52.5**

well-performing methods on Occluded-reID rely on the information from an external pose estimation model at inference [3, 5, 40, 28] or on occlusion data augmentation techniques [29, 31] to achieve robust part pooling on the new occluded domain. Different from these methods, we don't use any external model at inference, nor occlusion data augmentation, but still achieve competitive performance.

4.4. Ablation Studies

In this section, we conduct some ablations studies using the Occluded-Duke dataset and *BPBreID$_{RN}$*, to analyze the impact of our architectural choices on ReID performance.

4.4.1 Components of BPBreID

Performance gain related to different components of our model are reported in Table 2. We adopt Bag of Tricks (BoT) [14] as a baseline and build BPBreID on top of it.

BPBreID without learnable attention is an alternative method where the K body part probability maps $\{M_1, ..., M_K\}$ predicted by the body part attention module are replaced by fixed attention weights derived directly from PifPaf output, following a process detailed in the supplementary materials. The decreased performance primarily reveals that the lack of end-to-end training on the attention weights leads to a discrepancy between the fixed attention masks and the ReID need in terms of backbone feature pooling. This confirms that training the attention mechanism in an end-to-end fashion with both body part prediction and ReID as objectives leads to an attention mechanism which is more specialized towards solving the ReID task, with a better selection of discriminative appearance features.

BPBreID without visibility scores refers to a strategy where all embeddings are used at inference no matter their visibility, i.e., all visibility scores $v_i^{q|g}$ in Eq. (9) are set to 1. As expected, using noisy embeddings corresponding to non-visible parts dramatically reduces performance. This validates the effectiveness of our visibility-based part-to-part matching strategy for solving challenges $\langle 1 \rangle$ and $\langle 2 \rangle$.

Table 4. Impact of *identity loss* and *triplet loss* on training performance when applied selectively on holistic embeddings (*global* "g", *foreground* "f" and *concatenated* "c") and *body part-based embeddings* ("$p_{1,..,K}$"). Triplet loss on $p_{1,..,K}$ refers to our part-averaged triplet loss described in Section 3.3.2. Triplet loss on other embeddings (g, f and c) refers to a standard batch-hard triplet loss [8]. Identity loss on $p_{1,..,K}$ refers to a identity loss applied individually on each part-based embeddings. We also report performance for the popular part-based ReID architecture PCB [23], which does not use a foreground embedding.

Idx	Identity loss				Triplet loss				BPBreID (ours)		PCB [23]	
	g	f	c	$p_{1,..,K}$	g	f	c	$p_{1,..,K}$	R-1	mAP	R-1	mAP
GiLt	✓	✓	✓					✓	**66.7**	**54.1**	54.6	**46.3**
PCB				✓					57.2	43.2	51.2	40.8
1				✓	✓	✓	✓		52.9	43.2	50.2	40.9
2	✓	✓	✓	✓	✓	✓	✓	✓	59.5	48.2	51.1	42.8
3	✓	✓	✓	✓					61.5	49.5	52.1	44.8
4					✓	✓	✓	✓	53.9	41.9	45.5	37.6
5	✓	✓	✓	✓				✓	61.8	49.4	51.0	43.5
6	✓	✓						✓	65.5	51.4	52.9	43.5
7	✓		✓					✓	56.5	41.9	-	-
8		✓	✓					✓	64.0	52.9	**56.2**	46.2
9	✓	✓	✓				✓	✓	66.2	53.3	55.9	46.1
10	✓	✓	✓			✓		✓	63.6	52.2	-	-
11	✓	✓	✓		✓			✓	64.0	52.4	54.8	45.9
12	✓	✓	✓				✓		65.3	52.9	54.4	45.7

Our attempts to account for the visibility scores at training did not lead to performance improvement, but remains a promising path for future research. We speculate this happens because (1) GiLt is already robust to occlusion and (2) most training samples in Occluded-Duke are non occluded.

Finally, *BPBreID without part-averaged triplet loss* refers to a model where part-based embeddings are supervised individually with a classic triplet loss [8] instead of our part-averaged triplet loss described in Eq. (8). The difference between these two approaches lies in their underlying objective: the part-averaged triplet loss optimizes a global distance between two samples, which is computed using the average of local distances, whereas the standard triplet loss optimizes all local pairwise distances individually. The later approach gives reduced performance because it renders the training procedure more sensitive to occluded body-parts and non-discriminative local appearance. This last ablation test demonstrates the robustness of our part-averaged triplet loss to occlusions and non-discriminative local features, and therefore to solve challenges $\langle 1 \rangle$ and $\langle 3 \rangle$.

4.4.2 Validation of the GiLt Strategy

In this section, we study the impact of different combinations of the identity and triplet losses on the $K+3$ embeddings produced by our model. Results are reported in Table 4. We also report performance for the popular model architecture PCB [23], which partition the input image in six horizontal stripes, to demonstrate the superiority of our training scheme with other part-based architecture. The original PCB paper suggest a simple identity loss applied on part-based embeddings only: the corresponding sub-optimal performance is reported in the second table row. The experiment on the first row correspond to our *GiLt* strategy described in Section 3.3: holistic features are supervised only with an identity loss and part-based features are supervised with our part-averaged triplet loss. As demonstrated by experiments 1 to 4, triplet and identity losses are complementary to each other and best performance is reached when using them together. However, naively applying both losses on all embeddings (experiment 2) is a sub-optimal solution. We can draw two conclusions from experiments 5 to 8, which are small variations of our GiLt strategy regarding the identity loss. First, applying the identity loss on all three holistic embeddings leads to a more robust training scheme and to better performance. This experiment validates our choice of computing a global and a concatenated embeddings for training, even though we don't use them at inference. Second, experiment 5 validates our intuition that using an identity loss on part-based features is harmful to performance, since it renders the training procedure sensitive to occlusions and to non-discriminative local features. Experiments 9 to 12 validate our *GiLt* strategy of enforcing the triplet loss constraint on part-based embeddings only.

4.4.3 Discriminative Ability of Output Embeddings

In this Section, we study the discriminative ability of the holistic and body part-based embeddings $\{f_g, f_f, f_c, f_1, ..., f_K\}$ for $K = 6$. For that purpose, we compute query-to-gallery samples distance using each embedding individually or combination of them, and report the corresponding ranking performance in Table 3. When multiple embeddings are used, we compute the average distance weighted by visibility scores, as described in Eq. (9). As demonstrated in Table 3, performance using holistic embeddings are sub-optimal because the global embedding is sensitive to background clutter, and the foreground embedding cannot achieve part-to-part matching. The concatenated embedding fixes those issues but remains sensitive to occlusions because it contains noisy information from embeddings of non visible body parts. As demonstrated in the table, using body part-based embeddings individually leads to sub-optimal performance. Performance is better for embeddings from upper body parts, because (1) these parts are more discriminative and (2) lower body parts are very often occluded. Using all parts embeddings $\{f_1, ..., f_6\}$ leads to the best performance, because this strategy is the key to overcome the big challenges related to occluded re-id, i.e., (1) achieve feature alignment, (2) reduce background clutter and (3) compare

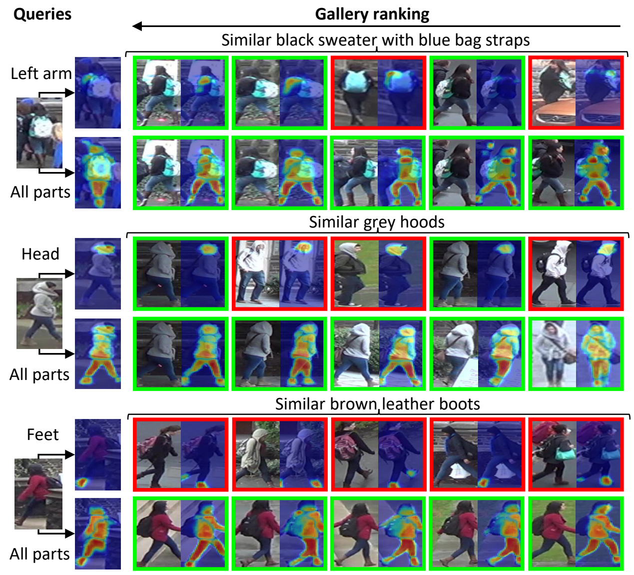

Figure 3. Visualization of ranking results based on individual body part-based embeddings (top row of each query) or all body part-based embeddings with the foreground embedding (bottom row of each query). For the "all parts" rows, only the foreground attention map is displayed for conciseness. In the top row of each query, the retrieved gallery samples are very similar w.r.t. the compared body part, but identities do not match because a single body part is not discriminative enough. Green/red borders are correct/incorrect matches. Best viewed in color and zoomed in.

only mutually visible body-parts. Finally, adding information from the foreground embedding produces slightly better performance, because it helps in mitigating errors caused by failed body part prediction, and by image pairs having few or no mutually visible parts. Figure 3 illustrates some ranking results using these embeddings individually.

5. Conclusions

In this work, we propose our model BPBreID to address the occluded person ReID task by learning body part representations and make two contributions. First, we design a body part attention module trained from a dual supervision with both identity and human parsing labels. With this attention mechanism, we show how external human semantic information can be effectively leveraged to produce ReID-relevant part-based features. Second, we investigate the influence of triplet and identity losses for learning part-based features and provide a simple yet effective GiLt strategy for training any part-based method. Our model achieves state-of-the-art performance on five popular ReID datasets.

Acknowledgment: This work has been funded by Sportradar, by the Walloon region project ReconnAIssance, and by the Fonds de la RechercheScientifique – FNRS.

References

[1] George Adaimi, Sven Kreiss, and Alexandre Alahi. Rethinking Person Re-Identification with Confidence. 6 2019.

[2] Peixian Chen, Wenfeng Liu, Pingyang Dai, Jianzhuang Liu, Qixiang Ye, Mingliang Xu, and Rongrong Ji. Occlude Them All : Occlusion-Aware Attention Network for Occluded Person Re-ID. *Iccv*, pages 11833–11842, 2021.

[3] Shang Gao, Jingya Wang, Huchuan Lu, and Zimo Liu. Pose-guided visible part matching for occluded person ReID. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 11741–11749, 2020.

[4] Silvio Giancola, Anthony Cioppa, Adrien Deliège, Floriane Magera, Vladimir Somers, Le Kang, Xin Zhou, Olivier Barnich, Christophe De Vleeschouwer, Alexandre Alahi, Bernard Ghanem, Marc Van Droogenbroeck, Abdulrahman Darwish, Adrien Maglo, Albert Clapés, Andreas Luyts, Andrei Boiarov, Artur Xarles, Astrid Orcesi, Avijit Shah, Baoyu Fan, Bharath Comandur, Chen Chen, Chen Zhang, Chen Zhao, Chengzhi Lin, Cheuk-Yiu Chan, Chun Chuen Hui, Dengjie Li, Fan Yang, Fan Liang, Fang Da, Feng Yan, Fufu Yu, Guanshuo Wang, H. Anthony Chan, He Zhu, Hongwei Kan, Jiaming Chu, Jianming Hu, Jianyang Gu, Jin Chen, João V. B. Soares, Jonas Theiner, Jorge De Corte, José Henrique Brito, Jun Zhang, Junjie Li, Junwei Liang, Leqi Shen, Lin Ma, Lingchi Chen, Miguel Santos Marques, Mike Azatov, Nikita Kasatkin, Ning Wang, Qiong Jia, Quoc Cuong Pham, Ralph Ewerth, Ran Song, Rengang Li, Rikke Gade, Ruben Debien, Runze Zhang, Sangrok Lee, Sergio Escalera, Shan Jiang, Shigeyuki Odashima, Shimin Chen, Shoichi Masui, Shouhong Ding, Sin-wai Chan, Siyu Chen, Tallal El-Shabrawy, Tao He, Thomas B. Moeslund, Wan-Chi Siu, Wei Zhang, Wei Li, Xiangwei Wang, Xiao Tan, Xiaochuan Li, Xiaolin Wei, Xiaoqing Ye, Xing Liu, Xinying Wang, Yandong Guo, Yaqian Zhao, Yi Yu, Yingying Li, Yue He, Yujie Zhong, Zhenhua Guo, and Zhiheng Li. SoccerNet 2022 Challenges Results. pages 75–86, 10 2022.

[5] Wang Guan'an, Shuo Yang, Huanyu Liu, Zhicheng Wang, Yang Yang, Shuliang Wang, Gang Yu, Erjin Zhou, and Jian Sun. High-order information matters: Learning relation and topology for occluded person re-identification. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 6448–6457, 2020.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2016-Decem, pages 770–778. IEEE Computer Society, 12 2016.

[7] Lingxiao He, Xingyu Liao, Wu Liu, Xinchen Liu, Peng Cheng, Tao Mei, and A I Research. FastReID: A Pytorch Toolbox for General Instance Re-identification. 6 2020.

[8] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In Defense of the Triplet Loss for Person Re-Identification. 3 2017.

[9] Houjing Huang, Xiaotang Chen, and Kaiqi Huang. Human Parsing Based Alignment With Multi-Task Learning For Occluded Person Re-Identification. pages 3–8, 2020.

[10] Mahdi M. Kalayeh, Emrah Basaran, Muhittin Gokmen, Mustafa E. Kamasak, and Mubarak Shah. Human Semantic Parsing for Person Re-identification. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1062–1071, 3 2018.

[11] Madhu Kiran, R Gnana Praveen, Le Thanh Nguyen-Meidine, Soufiane Belharbi, Louis-Antoine Blais-Morin, and Eric Granger. Holistic Guidance for Occluded Person Re-Identification. 4 2021.

[12] Sven Kreiss, Lorenzo Bertoni, and Alexandre Alahi. PifPaf: Composite fields for human pose estimation. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2019-June, pages 11969–11978. IEEE Computer Society, 3 2019.

[13] Yulin Li, Jianfeng He, Tianzhu Zhang, Xiang Liu, Yongdong Zhang, and Feng Wu. Diverse Part Discovery: Occluded Person Re-identification with Part-Aware Transformer. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, number 2, pages 2897–2906, 6 2021.

[14] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, volume 2019-June, pages 1487–1495, 2019.

[15] Jiaxu Miao, Yu Wu, Ping Liu, Yuhang DIng, and Yi Yang. Pose-guided feature alignment for occluded person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, volume 2019-Octob, pages 542–551, 2019.

[16] Xingang Pan, Ping Luo, Jianping Shi, and Xiaoou Tang. Two at Once: Enhancing Learning and Generalization Capacities via IBN-Net. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11208 LNCS:484–500, 7 2018.

[17] Yunjie Peng, Saihui Hou, Chunshui Cao, Xu Liu, Yongzhen Huang, and Zhiqiang He. Deep Learning-based Occluded Person Re-identification: A Survey. 7 2022.

[18] Guanqiu Qi, Gang Hu, Xiaofei Wang, Neal Mazur, Zhiqin Zhu, and Matthew Haner. EXAM: A Framework of Learning Extreme and Moderate Embeddings for Person Re-ID. *Journal of imaging*, 7(1), 1 2021.

[19] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C Berg, Li Fei-Fei, O Russakovsky, J Deng, H Su, J Krause, S Satheesh, S Ma, Z Huang, A Karpathy, A Khosla, M Bernstein, A C Berg, and L Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252, 9 2014.

[20] Chi Su, Jianing Li, Shiliang Zhang, Junliang Xing, Wen Gao, and Qi Tian. Pose-Driven Deep Convolutional Model for Person Re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, volume 2017-Octob, pages 3980–3989. Institute of Electrical and Electronics Engineers Inc., 9 2017.

[21] Yumin Suh, Jingdong Wang, Siyu Tang, Tao Mei, and Kyoung Mu Lee. Part-aligned bilinear representations for per-

son re-identification. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 11218 LNCS, pages 418–437. Springer Verlag, 4 2018.

[22] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2019-June, pages 5686–5696. IEEE Computer Society, 2 2019.

[23] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond Part Models: Person Retrieval with Refined Part Pooling (and a Strong Convolutional Baseline). *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11208 LNCS:501–518, 11 2017.

[24] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the Inception Architecture for Computer Vision. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2016-Decem, pages 2818–2826, 2016.

[25] Hongchen Tan, Xiuping Liu, Baocai Yin, and Xin Li. MHSA-Net: Multihead Self-Attention Network for Occluded Person Re-Identification. *IEEE Transactions on Neural Networks and Learning Systems*, 8 2022.

[26] Gabriel Van Zandycke, Vladimir Somers, Maxime Istasse, Carlo Del Don, and Davide Zambrano. DeepSportradar-v1: Computer Vision Dataset for Sports Understanding with High Quality Annotations. 22:1–8, 8 2022.

[27] Hong Xia Wang, Xiang Chen, and Chun Liu. Pose-guided part matching network via shrinking and reweighting for occluded person re-identification. *Image and Vision Computing*, 111:104186, 7 2021.

[28] Tao Wang, Hong Liu, Pinhao Song, Tianyu Guo, and Wei Shi. Pose-guided Feature Disentangling for Occluded Person Re-identification Based on Transformer. 2021.

[29] Zhikang Wang, Feng Zhu, Shixiang Tang, Rui Zhao, Lihuo He, and Jiangning Song. Feature Erasing and Diffusion Network for Occluded Person Re-Identification. 2021.

[30] Yunjie Xu, Liaoying Zhao, and Feiwei Qin. Dual attention-based method for occluded person re-identification. *Knowledge-Based Systems*, 212:106554, 1 2021.

[31] Cheng Yan, Guansong Pang, Jile Jiao, Xiao Bai, Xuetao Feng, and Chunhua Shen. Occluded Person Re-Identification with Single-scale Global Representations. In *ICCV 2021*, pages 11875–11884, 2021.

[32] Jinrui Yang, Jiawei Zhang, Fufu Yu, Xinyang Jiang, Mengdan Zhang, Xing Sun, Yingcong Chen, and Wei-Shi Zheng. Learning to Know Where to See : A Visibility-Aware Approach for Occluded Person Re-identification. *ICCV*, pages 11885–11894, 2021.

[33] Qin Yang, Peizhi Wang, Zihan Fang, and Qiyong Lu. Focus on the visible regions: Semantic-guided alignment model for occluded person re-identification. *Sensors (Switzerland)*, 20(16):1–15, 2020.

[34] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven C.H. Hoi. Deep Learning for Person Re-identification: A Survey and Outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1 2021.

[35] Xianghao Zang, Ge Li, Wei Gao, and Xiujun Shu. Learning to Disentangle Scenes for Person Re-identification. *Image and Vision Computing*, 116, 11 2021.

[36] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, and Zhibo Chen. Densely semantically aligned person re-identification. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2019-June, pages 667–676, 2019.

[37] Zhong Zhang, Haijia Zhang, and Shuang Liu. Person Re-identification using Heterogeneous Local Graph Attention Networks. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 12131–12140, 2021.

[38] Haiyu Zhao, Maoqing Tian, Shuyang Sun, Jing Shao, Junjie Yan, Shuai Yi, Xiaogang Wang, and Xiaoou Tang. Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, volume 2017-Janua, pages 907–915, 2017.

[39] Liming Zhao, Xi Li, Jingdong Wang, and Yueting Zhuang. Deeply-Learned Part-Aligned Representations for Person Re-Identification. *Proceedings of the IEEE International Conference on Computer Vision*, 2017-Octob:3239–3248, 7 2017.

[40] Kecheng Zheng, Cuiling Lan, Wenjun Zeng, Jiawei Liu, Zhizheng Zhang, and Zheng Jun Zha. Pose-Guided Feature Learning with Knowledge Distillation for Occluded Person Re-Identification. *MM 2021 - Proceedings of the 29th ACM International Conference on Multimedia*, 10:4537–4545, 7 2021.

[41] Liang Zheng, Yujia Huang, Huchuan Lu, and Yi Yang. Pose-Invariant Embedding for Deep Person Re-Identification. *IEEE Transactions on Image Processing*, 28(9):4500–4509, 1 2019.

[42] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE International Conference on Computer Vision*, volume 2015 Inter, pages 1116–1124, 2015.

[43] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled Samples Generated by GAN Improve the Person Re-identification Baseline in Vitro. In *Proceedings of the IEEE International Conference on Computer Vision*, volume 2017-Octob, pages 3774–3782. Institute of Electrical and Electronics Engineers Inc., 1 2017.

[44] Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. Re-ranking person re-identification with k-reciprocal encoding. In *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, volume 2017-Janua, pages 3652–3661. Institute of Electrical and Electronics Engineers Inc., 1 2017.

[45] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *AAAI 2020*

- *34th AAAI Conference on Artificial Intelligence*, pages 13001–13008. AAAI press, 8 2020.

[46] Kuan Zhu, Haiyun Guo, Zhiwei Liu, Ming Tang, and Jinqiao Wang. Identity-Guided Human Semantic Parsing for Person Re-Identification. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12348 LNCS:346–363, 7 2020.

[47] Kuan Zhu, Haiyun Guo, Shiliang Zhang, Yaowei Wang, Gaopan Huang, Honglin Qiao, Jing Liu, Jinqiao Wang, and Ming Tang. AAformer: Auto-Aligned Transformer for Person Re-Identification. 2021.

[48] Jiaxuan Zhuo, Zeyu Chen, Jianhuang Lai, and Guangcong Wang. Occluded Person Re-identification. *arXiv*, 4 2018.