Beyond Human Parts: Dual Part-Aligned Representations for Person Re-Identification

Jianyuan Guo[1†], Yuhui Yuan[2,3,4†], Lang Huang[1], Chao Zhang[1*], Jin-Ge Yao[2], Kai Han[5]
[1]Key Laboratory of Machine Perception (MOE), Peking University, [2]Microsoft Research Asia
[3]Institute of Computing Technology [4]University of Chinese Academy of Sciences [5]Noah's Ark Lab
{jyguo, laynehuang}@pku.edu.cn, chzhang@cis.pku.edu.cn
{yuhui.yuan, jinge.yao}@microsoft.com, kai.han@huawei.com

Abstract

Person re-identification is a challenging task due to various complex factors. Recent studies have attempted to integrate human parsing results or externally defined attributes to help capture human parts or important object regions. On the other hand, there still exist many useful contextual cues that do not fall into the scope of predefined human parts or attributes. In this paper, we address the missed contextual cues by exploiting both the accurate human parts and the coarse non-human parts. In our implementation, we apply a human parsing model to extract the binary human part masks and *a self-attention mechanism to capture the soft latent (non-human) part masks. We verify the effectiveness of our approach with new state-of-the-art performance on three challenging benchmarks: Market-*1501*, DukeMTMC-reID and CUHK*03*. Our implementation is available at https://github.com/ggjy/P2Net.pytorch.*

1. Introduction

Person re-identification has attracted increasing attention from both the academia and the industry in the past decade due to its significant role in video surveillance. Given an image for a particular person captured by one camera, the goal is to re-identify this person from images captured by different cameras from various viewpoints.

The task of person re-identification is inherently challenging because of the significant visual appearance changes caused by various factors such as human pose variations, lighting conditions, part occlusions, background cluttering and distinct camera viewpoints. All these factors make the misalignment problem become one of the most important problems for person re-identification task. With the surge of interest in deep representation learning, various approaches have been developed to address the misalignment problem, which could be roughly summarized as the following streams: (1) Hand-crafted partitioning, which relies on manually designed splits of the input image or the feature maps into grid cells [15, 38, 56] or horizontal stripes [1, 4, 41, 43, 51], based on the assumption that the human parts are well-aligned in the RGB color space. (2) The attention mechanism, which tries to learn an attention map over the last output feature map and constructs the aligned part features accordingly [55, 33, 50, 45]. (3) Predicting a set of predefined attributes [13, 37, 20, 2, 36] as useful features to guide the matching process. (4) Injecting human pose estimation [5, 11, 22, 35, 50, 54, 27] or human parsing results [10, 18, 34] to extract the human part aligned features based on the predicted human key points or semantic human part regions, while the success of such approaches heavily counts on the accuracy of human parsing models or pose estimators. Most of the previous studies mainly focus on learning more accurate human part representations, while neglecting the influence of potentially useful contextual cues that could be addressed as "non-human" parts.

Existing human parsing based approaches [50, 54] utilize an off-the-shelf semantic segmentation model to divide the input image into K predefined human parts, according to a predefined label set.[1] Beyond these predefined part categories, there still exist many objects or parts which could be critical for person re-identification, but tend to be recognized as background by the pre-trained human parsing models. For example, we illustrate some failure cases from human parsing results on the Market-1501 dataset in Figure 1. We can find that the objects belonging to undefined categories such as **backpack, reticule** and **umbrella** are in fact helpful and sometimes crucial for person re-identification. The existing human parsing datasets are mainly focused on parsing human regions, and most of these datasets fail to include all possible identifiable objects that could help person re-identification. Especially, most of the previous attention

[*] Corresponding author. [†] Equal contribution.

[1] E.g., the label set in [18]: background, hat, hair, glove, sunglasses, upper-clothes, dress, coat, socks, pants, jumpsuits, scarf, skirt, face, right-arm, left-arm, right-leg, left-leg, right-shoe and left-shoe.

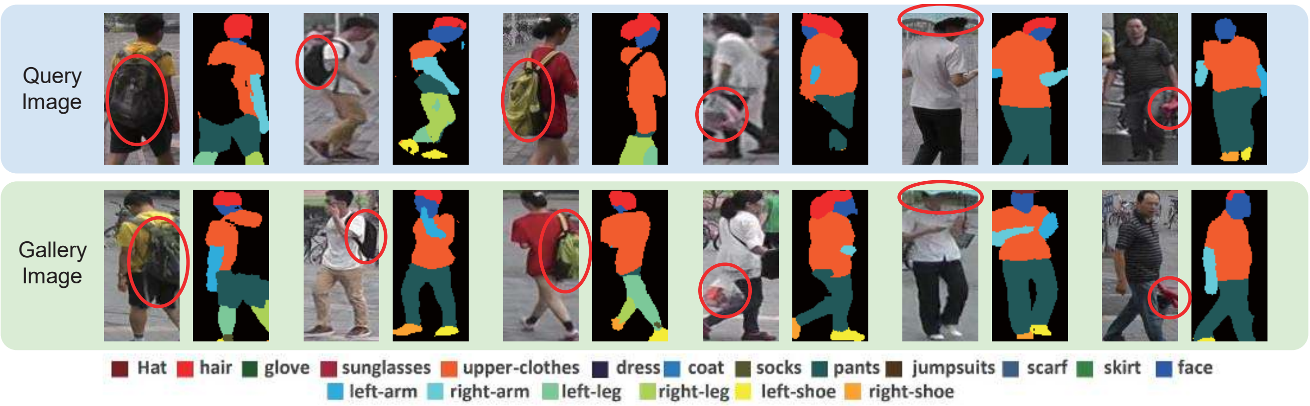

Figure 1: **Failure cases of the human parsing model**: The first row illustrates the query images, the second row illustrates the gallery images from Market-1501 and each column consists of two images belonging to the same identity. All of the regions marked with red circle are mis-classified as background (marked with black color) due to the limited label set while their ground-truth labels should be backpack, reticule and umbrella. It can be seen that these mis-classified regions are crucial for the person re-identification.

based approaches are mainly focused on extracting the human part attention maps.

Explicitly capturing useful information beyond predefined human parts or attributes has not been well studied in the previous literature. Inspired by the recently popular self-attention scheme [44, 48], we attempt to address the above problem by learning latent part masks from the raw data, according to the appearance similarities among pixels, which provide a coarse estimation of both human parts and the non-human parts, with the latter largely overlooked from the previous approaches based on human parsing.

Moreover, we propose the **dual part-aligned representation** scheme to combine the complementary information from both the accurate human parts and the coarse non-human parts. In our implementation, we apply a human parsing model to extract the human part masks and compute the human part-aligned representations for the features from the low-levels to high-levels. For the non-human part information, we apply the self-attention mechanism to learn to group all the pixels belonging to the same latent part together. We also extract the latent non-human part information on the feature maps from the low-levels to the high-levels. Through combining the advantages of both the accurate human part information and the coarse non-human part information, our approach learns to augment the representation of each pixel with the representation of the part (human parts *or* non-human parts) that it belongs to.

Our main contributions are summarized as below:

- We propose the **dual part-aligned representation** to update the representation by exploiting the complementary information from both the accurate human parts and the coarse non-human parts.
- We introduce the P^2-Net and show that our P^2-Net achieves new state-of-the-art performance on three benchmarks including Market-1501, DukeMTMC-reID and CUHK03.
- We analyze the contributions from both the human part representation and the latent part (non-human part) representation and discuss their complementary strengths in our ablation studies.

2. Related Work

The part misalignment problem is one of the key challenges for person re-identification, a host of methods [55, 35, 14, 54, 41, 27, 11, 10, 34, 49, 50, 38, 33, 8] have been proposed to mainly exploit the human parts to handle the body part misalignment problem, we briefly summarize the existing methods as below:

Hand-crafted Splitting for ReID. In previous studies, there are methods proposed to divide the input image or the feature map into small patches [1, 15, 38] or stripes [4, 43, 51] and then extract region features from the local patches or stripes. For instance, PCB [41] adopts a uniform partition and further refines every stripe with a novel mechanism. The hand-crafted approaches depend on the strong assumption that the spatial distributions of human bodies and human poses are exactly matching.

Semantic Segmentation for ReID. Different from the hand-crafted splitting approaches, [29, 35, 54, 10] apply a human part detector or a human parsing model to capture more accurate human parts. e.g., SPReID [10] utilizes a parsing model to generate 5 different predefined human part masks to compute more reliable part representations, which achieves promising results on various person re-identification benchmarks.

Poses/Keypoints for ReID. Similar to the semantic segmentation approaches, poses or keypoints estimation can also be used for accurate/reliable human part localization. For example, there are approaches exploring both the human poses and the human part masks [9], or generating human part masks via exploting the connectivity of the keypoints [50]. There are some other studies [5, 29, 35, 54] that also exploit the pose cues to extract the part-aligned features.

Attention for ReID. Attention mechanisms have been used to capture human part information in recent work [21, 55, 50, 17, 34]. Typically, the predicted attention maps distribute most of the attention weights on human parts that may help improve the results. To the best of our knowledge, we find that most of the previous attention approaches are limited to capturing the human part only.

Attributes for ReID. Semantic attributes [46, 25, 7] have been exploited as feature representations for person re-identification tasks. Previous work [47, 6, 20, 42, 57] leverages the attribute labels provided by original dataset to generate attribute-aware feature representation. Different from previous work, our latent part branch can attend to important visual cues without relying on detailed supervision signals from the limited predefined attributes.

Our Approach. To the best of our knowledge, we are the first to explore and define the (non-human) contextual cues. We empirically demonstrate the effectiveness of combining separately crafted components for the well-defined, accurate human parts *and* all other potentially useful (but coarse) contextual regions.

3. Approach

First, we present our key contribution: **dual part-aligned representation**, which learns to combine both the accurate human part information and the coarse latent part information to augment the representation of each pixel (Sec. 3.1). Second, we present the network architecture and the detailed implementation of P^2-Net (Sec. 3.2).

3.1. Dual Part-Aligned Representation

Our approach consists of two branches: a *human part branch* and a *latent part branch*. Given an input feature map $\mathbf{X}$ of size $N \times C$, where $N = H \times W$, H and W are the height and width of the feature map, C is the number of channels, we apply the human part branch to extract accurate human part masks and compute the human part-aligned representation $\mathbf{X}^{\text{Human}}$ accordingly. We also use a latent part branch to learn to capture both the coarse non-human part masks and the coarse human part masks based on the appearance similarities between different pixels, then we compute the latent part-aligned representation $\mathbf{X}^{\text{Latent}}$ according to the coarse part masks. Last, we augment the original representation with both the human part-aligned representation and the latent part-aligned representation.

Human Part-Aligned Representation. The main idea of the human part-aligned representation is to represent each pixel with the human part representation that the pixel belongs to, which is the aggregation of the pixel-wise representations weighted by a set of confidence maps. Each confidence map is used to surrogate a semantic human part.

We illustrate how to compute the human part-aligned representation in this section. Assuming there are $K-1$ predefined human part categories in total from a human parsing model, we treat all the rest proportion of regions in the image as the background according to the human parsing result. In summary, we need to estimate K confidence maps for the human part branch.

We apply the state-of-the-art human parsing framework CE2P [23] to predict the semantic human part masks for all the images in all three benchmarks in advance, as shown in Figure 2(b). We denote the predicted label map of image $\mathbf{I}$ as $\mathbf{L}$. We re-scale the label map $\mathbf{L}$ to be of the same size as the feature map $\mathbf{X}$ ($\mathbf{x}_i$ is the representation of pixel i, essentially the i_{th} row of $\mathbf{X}$) before using it. We use l_i to represent the human part category of pixel i within the re-scaled label map, and l_i is of K different values including $K-1$ human part categories and one background category.

We denote the K confidence maps as $\mathbf{P}_1, \mathbf{P}_2, \cdots, \mathbf{P}_K$, where each confidence map $\mathbf{P}_k$ is associated with a human part category (or the background category). According to the predicted label map $\mathbf{L}$, we set $p_{ki} = 1$ (p_{ki} is the i_{th} element of $\mathbf{P}_k$) if $l_i \equiv k$ and $p_{ki} = 0$ otherwise. Then we apply L1 normalization on each confidence map and compute the human part representation as below,

$$\mathbf{h}_k = g(\sum_{i=1}^{N} p_{ki}\mathbf{x}_i), \tag{1}$$

where $\mathbf{h}_k$ is the representation of the k_{th} human part, g function is used to learn better representation and p_{ki} is the confidence score after L1 normalization. Then we generate the human part-aligned feature map $\mathbf{X}^{\text{Human}}$ of the same size as the input feature map $\mathbf{X}$, and each element of $\mathbf{X}^{\text{Human}}$ is set as

$$\mathbf{x}_i^{\text{Human}} = \sum_{k=1}^{K} \mathbb{1}[l_i \equiv k]\mathbf{h}_k, \tag{2}$$

where $\mathbb{1}[l_i \equiv k]$ is an indicator function and each $\mathbf{x}_i^{\text{Human}}$ is essentially the part representation of the semantic human part that it belongs to. For the pixels predicted as the background, we choose to aggregate the representations of all pixels that are predicted as the background and use it to augment their original representations.

Latent Part-Aligned Representation. We explain how to estimate the latent part representation in this section. Since we can not predict accurate masks for non-human cues based on the existing approaches, we adapt the self-attention mechanism [44, 48] to enhance our framework by learning to capture some coarse latent parts automatically from data based on the semantic similarities between each pixel and all other pixels. The latent part is expected to capture details that are weakly utilized in the human part

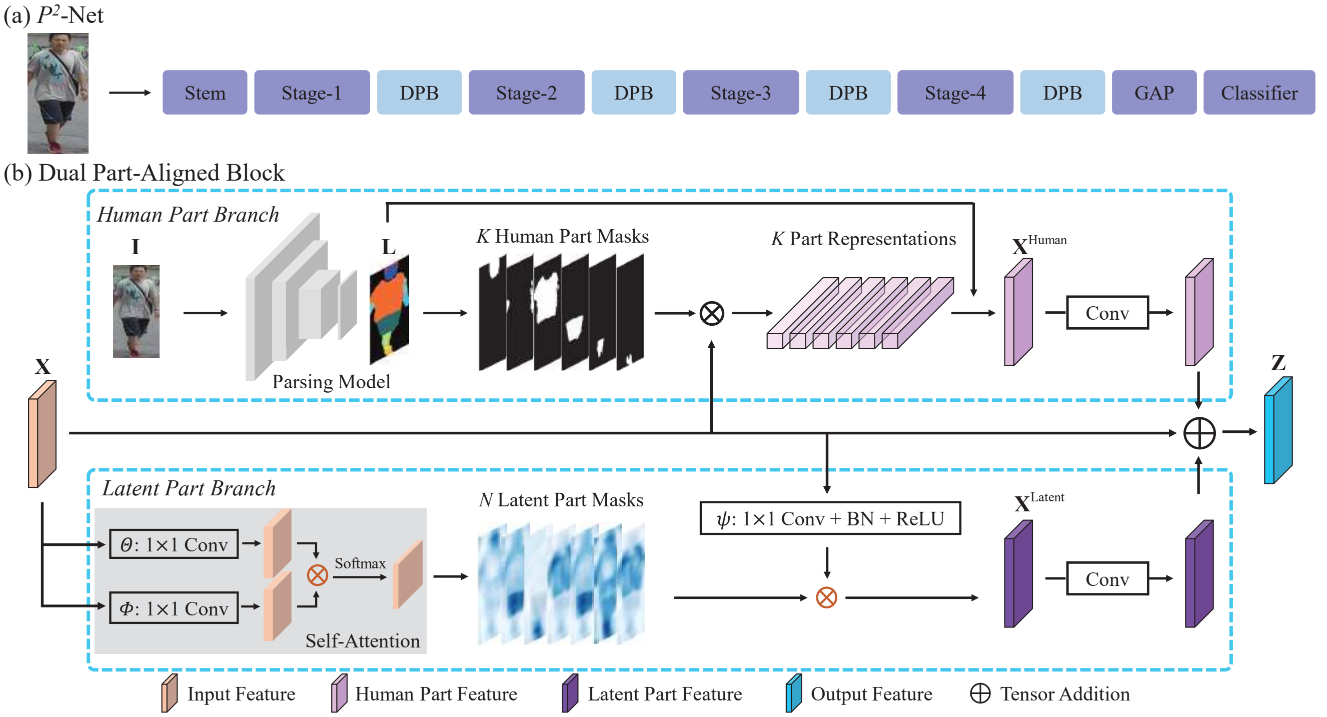

Figure 2: Illustration of the overall structure of P^2-Net and the dual part-aligned block (DPB). (a) Given an input image, we employ a ResNet-50 backbone consists of a stem, four stages (e.g., Res-1, Res-2, Res-3 and Res-4), global average pooling (GAP) and a classifier. We insert a DPB after every stage within the ResNet backbone. (b) The DPB consists of a human part branch and a latent part branch. For the human part branch, we employ the CE2P [23] to predict the human part label maps and generate the human part masks accordingly. For the latent part branch, we employ the self-attention scheme to predict the latent part masks. We compute the human part-aligned representation and latent part-aligned representation within the two branches separately. Last, we add the outputs from these two branches to the input feature map as the final output feature map.

branch. We are particularly interested in the contribution from the coarse non-human part masks on the important cues that are missed by the predefined human parts or attributes.

In our implementation, the latent part branch learns to predict N coarse confidence maps $\mathbf{Q}_1, \mathbf{Q}_2, \cdots, \mathbf{Q}_N$ for all N pixels, each confidence map $\mathbf{Q}_i$ learns to pay more attention to the pixels that belong to the same latent part category as the i_{th} pixel.

We illustrate how to compute the confidence map for the pixel i as below,

$$q_{ij} = \frac{1}{Z_i} \exp(\theta(\mathbf{x}_j)^\top \phi(\mathbf{x}_i)), \tag{3}$$

where q_{ij} is the j_{th} element of $\mathbf{Q}_i$, $\mathbf{x}_i$ and $\mathbf{x}_j$ are the representations of the pixels i and j respectively. $\theta(\cdot)$ and $\phi(\cdot)$ are two transform functions to learn better similarities and are implemented as 1×1 convolution, following the self-attention mechanism [44, 48]. The normalization factor Z_i is a sum of all the similarities associated with pixel i: $Z_i = \sum_{j=1}^{N} \exp(\theta(\mathbf{x}_j)^\top \phi(\mathbf{x}_i))$.

Then we estimate the latent part-aligned feature map $\mathbf{X}^{\text{Latent}}$ as below,

$$\mathbf{x}_i^{\text{Latent}} = \sum_{j=1}^{N} q_{ij} \psi(\mathbf{x}_j), \tag{4}$$

where $\mathbf{x}_i^{\text{Latent}}$ is the i_{th} element of $\mathbf{X}^{\text{Latent}}$. We estimate the latent part-aligned representation for pixel i by aggregating the representations of all the other pixels according to their similarities with pixel i. ψ is a function used to learn better representation, which is implemented with 1×1 convolution + BN + ReLU.

For the latent part-aligned representation, we expect each pixel can pay more attention to the part that it belongs to, which is similar with the recent work [12, 53]. The self-attention is a suitable mechanism to group the pixels with similar appearance together. We empirically study the influence of the coarse human part information and the coarse non-human part information to verify the effectiveness is mainly attributed to the coarse non-human parts (Sec. 4.3).

Last, we fuse the human part-aligned representation and the latent part-aligned representation as below,

$$\mathbf{Z} = \mathbf{X} + \mathbf{X}^{\text{Human}} + \mathbf{X}^{\text{Latent}}, \tag{5}$$

where $\mathbf{Z}$ is the final representation of our approach.

3.2. P^2-Net

Backbone. We use ResNet-50 pre-trained on the ImageNet as the backbone following the previous PCB [41].

Dual Part-Aligned Representation. In our implementation, we employ the dual part-aligned block (DPB) after Res-1, Res-2, Res-3 and Res-4 stages. Assuming that the input image is of size 384×128, the output feature map from Res-1/Res-2/Res-3/Res-4 stage is of size 96×32/48×16/ 24×8/24×8 respectively. We have conducted detailed ablation study about DPB in Section 4.3. For the human part branch, we employ the CE2P [23] model to extract the human part label maps of size 128×64, then we resize the label maps to be of the size 96×32/48×16/24×8/24×8 for the four stages respectively. For the latent part branch, we employ the self-attention mechanism on the output feature map from each stage directly.

Network Architecture. The ResNet backbone takes an image $\mathbf{I}$ as input and outputs feature map $\mathbf{X}$ after the Res-4 stage. We feed the feature map $\mathbf{X}$ into the global average pooling layer and employ the classifier at last. We insert the DPB after every stage to update the representations before feeding the feature map into the next stage. We could achieve better performance through applying more DPBs. The overall pipeline is illustrated in Figure 2(a).

Loss Function. All of our baseline experiments only employ the softmax loss to ensure the fairness of the comparison and for ease of ablation study. To compare with the state-of-the-art approaches, we further employ the triplet loss following the previous work.

4. Experiments

4.1. Datasets and Metrics

Market-1501. Market-1501 dataset [58] consists of 1501 identities captured by 6 cameras, where the train set consists of $12,936$ images of 751 identities, the test set is divided into a query set that contains $3,368$ images and a gallery set that contains $16,364$ images.

DukeMTMC-reID. DukeMTMC-reID dataset [28, 59] consists of $36,411$ images of $1,404$ identities captured by 8 cameras, where the train set contains $16,522$ images, the query set consists of $2,228$ images and the gallery set consists of $17,661$ images.

CUHK03. CUHK03 dataset [15] contains $14,096$ images of $1,467$ identities captured by 6 cameras. CUHK03 provides two types of data, hand-labeled ("labeled") and DPM-detected ("detected") bounding boxes, the latter type is more challenging due to severe bounding box misalignment and cluttered background. We conduct experiments on both "labeled" and "detected" types of data. We split the dataset following the training/testing split protocol proposed in [60], where the train/query/gallery set consists of $7,368$/$1,400$/$5,328$ images respectively.

We employ two kinds of evaluation metrics including the cumulative matching characteristics (CMC) and mean average precision (mAP). Especially, all of our experiments employ the single-query setting without any other post-processing techniques such as re-ranking [60].

4.2. Implementation Details

We choose ResNet-50 pre-trained on ImageNet as our backbone. After getting the feature map from the last residual block, we use a global average pooling and a linear layer (FC+BN+ReLU) to compute a 256-D feature embedding. We use ResNet-50 trained with softmax loss as our baseline model, and set the stride of the last stage in ResNet from 2 to 1 following [41]. We also use triplet loss [4, 19, 55] to improve the performance.

We use the state-of-the-art human parsing model CE2P [23] to predict the human part label maps for all the images in the three benchmark in advance. The CE2P model is trained on the Look Into Person [18] (LIP) dataset, which consists of $\sim$$30,000$ finely annotated images with 20 semantic labels (19 human parts and 1 background). We divide the 20 semantic categories into K groups [2], and train the CE2P model with the grouped labels. We adopt the training strategies as described in CE2P [23].

All of our implementations are based on PyTorch framework [26]. We resize all the training images to 384×128 and then augment them by horizontal flip and random erasing [61]. We set the batch size as 64 and train the model with base learning rate starts from 0.05 and decays to 0.005 after 40 epochs, the training is finished at 60 epochs. We set momentum $\mu = 0.9$ and the weight decay as 0.0005. All of the experiments are conducted on a single NVIDIA TITAN XP GPU.

4.3. Ablation study

The core idea of DPB lies on the human part branch and the latent part branch. We perform comprehensive ablation studies of them in follows.

Influence of the part numbers for human part branch. As we can divide the input image into different number of parts in different levels. we study the impact of the number of different semantic parts (i.e., K = 1, K = 2, K = 5) on the Market-1501 benchmark. We summarize all of the results in Table 1. The 1^{st} row reports the results of baseline model and the the 2^{nd} row to 4^{th} report the performances that only apply the human part branch with different choices of K. When $K = 1$, there is no extra parsing information added to the network and the performances keep almost the

[2] When K = 5, each group represents background, head, upper-torso, lower-torso and shoe; when K = 2, it represents background and foreground; when $K = 1$, it treats the whole image as a single part.

Table 1: Ablation study of the DPB on Market-1501. K is the number of human parts within the human part branch, We insert the DPB after the stage-k (Res-k), where $k = 1, 2, 3, 4$. We employ HP-p to represent the human part branch choosing $K = p$. DPB (HP-p) represents using the human part branch only while DPB (Latent) represents using the latent part branch only.

Method	Res-1			Res-2			Res-3			Res-4		
	R-1	R-5	mAP	R-1	R-5	mAP	R-1	R-5	mAP	R-1	R-5	mAP
Baseline	88.36	95.39	71.48	88.36	95.39	71.48	88.36	95.39	71.48	88.36	95.39	71.48
DPB (HP-1)	88.98	95.22	71.37	90.12	96.08	74.20	89.62	95.67	72.82	89.51	95.55	72.19
DPB (HP-2)	90.17	96.35	74.49	90.63	96.67	75.87	90.74	96.39	76.74	90.22	96.34	74.11
DPB (HP-5)	90.77	96.44	77.22	91.83	96.89	78.72	91.23	96.74	77.21	90.26	96.29	75.46
DPB (Latent)	90.20	96.40	73.28	91.73	96.86	78.48	91.47	96.86	77.80	89.31	96.14	73.71
DPB (HP-5 + Latent)	**91.00**	**96.88**	**76.99**	**92.75**	**97.45**	**80.98**	**91.87**	**97.13**	**78.80**	**91.18**	**97.03**	**78.36**

Table 2: Ablation study of the human-part (Latent w/o NHP) and non-human part (Latent w/o HP) in the latent part branch.

HP-5	Latent w/o NHP	Latent w/o HP	Market Res-2		Market Res-3	
			R-1	mAP	R-1	mAP
-	-	-	88.36	71.48	88.36	71.48
×	✓	×	91.19	77.22	91.12	77.10
×	×	✓	91.55	78.25	91.35	77.23
×	✓	✓	**91.73**	**78.48**	**91.47**	**77.80**
			Market Res-2		CUHK (detected)	
✓	×	×	91.83	78.72	67.57	60.02
✓	✓	×	91.97	79.31	68.46	61.98
✓	×	✓	**92.56**	**80.60**	**69.61**	**62.85**

Table 3: Comparison of using 1, 3 and 5 DPBs on the Market-1501. DPB consists of both the human part branch and the latent part branch here.

Method	HP-5	Latent	R-1	R-5	R-10	mAP
Baseline	-	-	88.36	95.39	97.06	71.48
+ 1 × DPB	✓	×	91.83	96.89	97.95	78.72
+ 3 × DPB	✓	×	92.01	97.15	98.16	78.87
+ 5 × DPB	✓	×	92.26	97.26	98.20	79.28
+ 1 × DPB	×	✓	91.73	96.86	98.10	78.48
+ 3 × DPB	×	✓	92.12	97.32	98.28	80.15
+ 5 × DPB	×	✓	92.79	97.65	98.52	80.49
+ 1 × DPB	✓	✓	92.75	97.45	98.22	80.98
+ 3 × DPB	✓	✓	93.28	97.79	98.61	82.08
+ 5 × DPB	✓	✓	**93.96**	**97.98**	**98.81**	**83.40**

Table 4: Comparison of the two branches of DPB on CUHK03.

Method	HP-5	Latent	R-1	R-5	R-10	mAP
Baseline	-	-	60.29	78.21	84.86	54.79
+ 5 × DPB	✓	×	69.93	83.86	88.90	63.34
	×	✓	69.84	83.50	89.83	63.25
	✓	✓	**71.55**	**85.71**	**90.80**	**64.23**

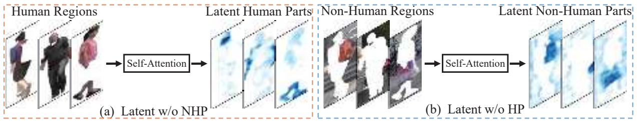

Figure 3: Latent w/o NHP vs. Latent w/o HP: *Latent w/o NHP* only applies self-attention on the human part regions while *Latent w/o HP* only applies self-attention on the non-human part regions. The human/non-human part regions are based on the human parsing prediction.

same with the baseline model. When $K = 2$, the human part branch introduces the foreground and the background contextual information to help extracting more reliable human context information. we can observe obvious improvements in R-1 and mAP compared to the previous two results. The performance improves with larger K, which indicates that accurately aggregating contextual information from pixels belonging to same semantic human part is crucial for person re-identification. We set $K = 5$ as the default setting for human part branch if not specified.

Non-human part in latent part branch. The choice of self-attention for latent part branch is mainly inspired by that self-attention can learn to group the similar pixels together without extra supervision (also shown useful in segmentation [53, 12]). Considering that the latent part branch is in fact the mixture of the coarse human and non-human part information, we empirically verify that the performance gains from the latent part branch is mainly attributed to capturing non-human parts, as shown in Table 2. We use the binary masks predicted by the human parsing model (K = 2) to control the influence of the human regions or non-human regions within the latent part branch. Here we study two kinds of settings: (1) only use the non-human part information within the latent part branch, we apply the the binary human masks (1 for the non-human pixels and 0 for the human pixels) to remove the influence of the pixels predicted as human parts, called as Latent w/o HP. (2) only use the human part information within the latent part branch, we also apply the binary human masks (1 for the human pixels and 0 for the non-human pixels) to remove the influence of the pixels predicted as non-human parts, called as Latent w/o NHP. It can be seen that the gains of the latent part branch mainly comes from the help of non-human part information, the Latent w/o HP outperforms the Latent w/o NHP and is very close to the original latent part branch.

Besides, we also study the contribution of the latent branch when applying the human part branch (HP-5). We choose the DPB (HP-5) inserted after Res-2 as our baseline and add the latent part branch that applies self-attention on either the human regions only (Latent w/o NHP in Fig. 3) or non-human regions only (Latent w/o HP in Fig. 3). It can be seen that DPB (HP-5 + Latent w/o HP) largely outperforms DPB (HP-5 + Latent w/o NHP) and is close to DPB (HP-5 + Latent), which further verifies the effectiveness of the latent part branch is mainly attributed to exploiting the non-human parts.

Complementarity of two branches. Dual part-aligned block (DPB) consists of the human part branch and the latent part branch. The human part branch helps improving the performance by eliminating the influence of the noisy background context information, and the latent part branch introduces the latent part masks to surrogate various non-

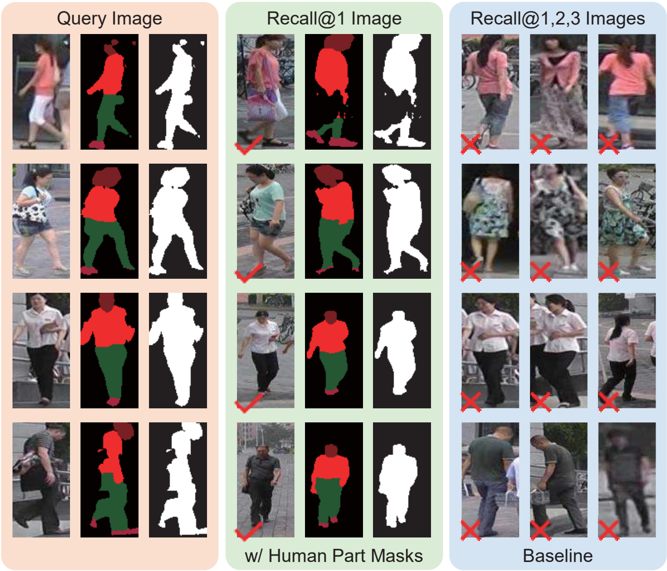

Figure 4: Comparison of Baseline and P^2-Net that only employs the human part branch. For all of the four query images, Recall@3 of the Baseline method is 0, while the Recall@1 of the P^2-Net (w/ Human Part Masks) is 1. The 1^{st} and 2^{nd} rows illustrate the cases that the bag is visible in one viewpoint but invisible in other viewpoints, human part masks eliminate the influence of the bags as the bags are categorized as background. The 3^{rd} and 4^{th} rows illustrate cases that the area of person only occupies small proportions of the whole images and the background context information leads to poor performance, human part masks can eliminate the influence of background regions.

human parts.

We empirically show that the two branches are complementary with the experimental results on the 6^{th} row of Table 1. It can be seen that combining both the human part-aligned representation and the latent part-aligned representation boosts the performance for all stages. We can draw the following conclusions from Table 3 and Table 4: (i) although the latent part masks are learned from scratch, DPB (latent) achieves comparable results with the human part branch in general, which carries more strong prior information of the human parts knowledge, showing the importance of the non-human part context. (ii) human part branch and latent part branch are complementary to each other. In comparison to the results only using a single branch, inserting $5\times$ DPB attains 1% and 3% gain in terms of R-1 and mAP on Market-1501, 1.6% and 1% gain in terms of R-1 and mAP on CUHK03, respectively.

We visualize the predicted human part masks to illustrate how it helps improving the performance in Figure 4. For all of the four query images, the baseline method fails to return the correct images of the same identity while we can find the correct images by employing the human part masks. In summary, we can see that the context information of the non-informative background influences the final results and the human part masks eliminate the influence of these noisy context information.

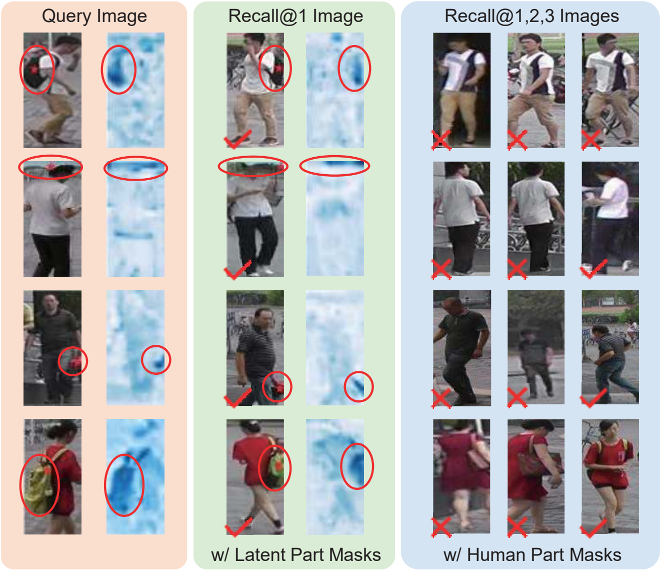

Figure 5: Comparison of P^2-Net (w/ Latent Part Masks) and P^2-Net (w/ Human Part Masks). There exist some important non-human parts in all of the four query images. The P^2-Net (w/ Human Part Masks) categorizes these crucial non-human parts as background and fails to return the correct image at Recall@1. The P^2-Net (w/ Latent Part Masks) predicts the latent part mask associated with these non-human parts, which successfully returns the correct image at Recall@1. It can be seen that the predicted latent part masks serves as reliable surrogate for the non-human part.

There also exist large amounts of scenarios that the non-human part context information is the key factor. We illustrate some typical examples in Figure 5, and we mark the non-human but informative parts with the red circles. For example, the 1^{st} and 4^{th} row illustrate that mis-classifying the **bag** as background causes the failures of the human part masks based method. Our approach addresses these failure cases through learning the latent part masks and it can be seen that the predicted latent part masks within the latent part branch well surrogate the non-human but informative parts. In summary, the human part branch benefits from the latent part branch through dealing with the non-human part information.

Number of DPB. To study the influence of the numbers of DPB (with human part representation only, with latent part representation only and with both human and latent part representations), we add 1 block (to Res-2), 3 blocks (2 to Res-2, and 1 to Res-3) and 5 blocks (2 to Res-2, and 3 to Res-3) within the backbone network. As shown in Table 3, more DPB blocks lead to better performance. We achieve the best performance with 5 DPBs, which boosts the R-1 accuracy and mAP by 5.6% and 11.9% respectively. We set the number of DPB block as 5 in all of our state-of-the-art experiments.

Table 5: Comparison with the SOTA on Market-1501.

Method	R-1	R-5	R-10	mAP
Spindle [54]	76.9	91.5	94.6	-
MGCAM [34]	83.8	-	-	74.3
PDC [35]	84.1	92.7	94.9	63.4
AACN [50]	85.9	-	-	66.9
PSE [29]	87.7	94.5	96.8	69.0
PABR [39]	90.2	96.1	97.4	76.0
SPReID [10]	92.5	97.2	98.1	81.3
MSCAN [14]	80.3	-	-	57.5
DLPAR [55]	81.0	92.0	94.7	63.4
SVDNet [40]	82.3	92.3	95.2	62.1
DaF [52]	82.3	-	-	72.4
JLML [16]	85.1	-	-	65.5
DPFL [3]	88.9	-	-	73.1
HA-CNN [17]	91.2	-	-	75.7
SGGNN [32]	92.3	96.1	97.4	82.8
GSRW [31]	92.7	96.9	98.1	82.5
PCB + RPP [41]	93.8	97.5	98.5	81.6
P^2-Net	94.0	98.0	98.8	83.4
P^2-Net (+ triplet loss)	**95.2**	**98.2**	**99.1**	**85.6**

4.4. Comparison with state-of-the-art

We empirically verify the effectiveness of our approach with a series of state-of-the-art (SOTA) results on all of the three benchmarks. We illustrate more details as following.

We illustrate the comparisons of our P^2-Net with the previous state-of-the-art methods on Market-1501 in Table 5. Our P^2-Net outperforms all the previous methods by a large margin. We achieve a new SOTA performance such as R-1=95.2% and mAP=85.6% respectively. Especially, our P^2-Net outperforms the previous PCB by 1.8% in mAP without using multiple softmax losses for training. When equiped with the triplet loss, our P^2-Net still outperforms the PCB by 1.4% and 4.0% in terms of R-1 and mAP, respectively. Besides, our proposed P^2-Net also outperforms the SPReID [10] by 2.7% measured by R-1 accuracy.

We summarize the comparisons on DukeMTMC-reID in Table 6. It can be seen that P^2-Net surpasses all the previous SOTA methods. SPReID [10] is the method has the closest performance with us in R-1 accuracy. Notably, the SPReID train their model with more than 10 extra datasets to improve the performance while we only use the original dataset as the training set.

Last, we evaluate our P^2-Net on CUHK03 dataset. We follow the training/testing protocol proposed by [60]. As illustrated in Table 7, our P^2-Net outperforms the previous SOTA method MGCAM [34] by 28.2% measured by R-1 accuracy and 23.4% measured by mAP. For the CUHK03-detected dataset, our P^2-Net still outperforms the previous SOTA method PCB+RPP [41] by 11.2% measured by R-1 accuracy and 11.4% measured by mAP.

Table 6: Comparison with the SOTA on DukeMTMC-reID.

Method	R-1	R-5	R-10	mAP
AACN [50]	76.8	-	-	59.3
PSE [29]	79.8	89.7	92.2	62.0
PABR [39]	82.1	90.2	92.7	64.2
SPReID [10]	84.4	91.9	93.7	71.0
SBAL [24]	71.3	-	-	52.4
ACRN [30]	72.6	84.8	88.9	52.0
SVDNet [40]	76.7	86.4	89.9	56.8
DPFL [3]	79.2	-	-	60.6
SVDEra [60]	79.3	-	-	62.4
HA-CNN [17]	80.5	-	-	63.8
GSRW [31]	80.7	88.5	90.8	66.4
SGGNN [32]	81.1	88.4	91.2	68.2
PCB + RPP [41]	83.3	90.5	92.5	69.2
P^2-Net	84.9	92.1	94.5	70.8
P^2-Net (+ triplet loss)	**86.5**	**93.1**	**95.0**	**73.1**

Table 7: Comparison with the SOTA on CUHK03.

Method	labeled		detected	
	R-1	mAP	R-1	mAP
DaF [52]	27.5	31.5	26.4	30.0
SVDNet [40]	40.9	37.8	41.5	37.3
DPFL [3]	43.0	40.5	40.7	37.0
HA-CNN [17]	44.4	41.0	41.7	38.6
SVDEra [60]	49.4	45.1	48.7	43.5
MGCAM [34]	50.1	50.2	46.7	46.9
PCB + RPP [41]	-	-	63.7	57.5
P^2-Net	75.8	69.2	71.6	64.2
P^2-Net (+ triplet loss)	**78.3**	**73.6**	**74.9**	**68.9**

In summary, our proposed P^2-Net outperforms all the previous approaches by a large margin and achieves new state-of-the-art performance on all the three challenging benchmarks.

5. Conclusion

In this work, we propose a novel **dual part-aligned representation** scheme to address the non-human part misalignment problem for person re-identification. It consists of a human part branch and a latent part branch to tackle both human part misalignment and non-human part misalignment problem. The human part branch adopts off-the-shelf human parsing model to inject structural prior information by capturing the predefined semantic human parts for a person, while the latent part branch adopts a self-attention mechanism to help capture the detailed part categories beyond the injected prior information. Based on dual part-aligned representation, our approach achieves the state-of-the-art performances on all of the three benchmarks including Market-1501, DukeMTMC-reID and CUHK03.

Acknowledgement

This work was supported by the National Natural Science Foundation of China under Grant 61671027 and the National Key Basic Research Program of China under Grant 2015CB352303.

References

[1] Ejaz Ahmed, Michael Jones, and Tim K Marks. An improved deep learning architecture for person re-identification. In *CVPR*, 2015.

[2] Xiaobin Chang, Timothy M Hospedales, and Tao Xiang. Multi-level factorisation net for person re-identification. In *CVPR*, 2018.

[3] Yanbei Chen, Xiatian Zhu, Shaogang Gong, et al. Person re-identification by deep learning multi-scale representations. In *ICCV*, 2018.

[4] De Cheng, Yihong Gong, Sanping Zhou, Jinjun Wang, and Nanning Zheng. Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In *CVPR*, 2016.

[5] Yeong-Jun Cho and Kuk-Jin Yoon. Improving person re-identification via pose-aware multi-shot matching. In *CVPR*, 2016.

[6] Kai Han, Jianyuan Guo, Chao Zhang, and Mingjian Zhu. Attribute-aware attention model for fine-grained representation learning. In *ACMMM*, 2018.

[7] Kai Han, Yunhe Wang, Han Shu, Chuanjian Liu, Chunjing Xu, and Chang Xu. Attribute aware pooling for pedestrian attribute recognition. *arXiv:1907.11837*, 2019.

[8] Lingxiao He, Jian Liang, Haiqing Li, and Zhenan Sun. Deep spatial feature reconstruction for partial person re-identification: Alignment-free approach. In *CVPR*, 2018.

[9] Dong Jian, Chen Qiang, Xiaohui Shen, Jianchao Yang, and Shuicheng Yan. Towards unified human parsing and pose estimation. In *CVPR*, 2014.

[10] Mahdi M Kalayeh, Emrah Basaran, Muhittin Gökmen, Mustafa E Kamasak, and Mubarak Shah. Human semantic parsing for person re-identification. In *CVPR*, 2018.

[11] Vijay Kumar, Anoop Namboodiri, Manohar Paluri, and C. V. Jawahar. Pose-aware person recognition. In *CVPR*, 2017.

[12] Huang Lang, Yuan Yuhui, Guo Jianyuan, Zhang Chao, Chen Xilin, and Wang Jingdong. Interlaced sparse self-attention for semantic segmentation. *arXiv:1907.12273*, 2019.

[13] Ryan Layne, Timothy M Hospedales, and Shaogang Gong. Attributes-based re-identification. In *Person Re-Identification*. 2014.

[14] Dangwei Li, Xiaotang Chen, Zhang Zhang, and Kaiqi Huang. Learning deep context-aware features over body and latent parts for person re-identification. In *CVPR*, 2017.

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

[16] Wei Li, Xiatian Zhu, and Shaogang Gong. Person re-identification by deep joint learning of multi-loss classification. In *IJCAI*, 2017.

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

[18] Xiaodan Liang, Ke Gong, Xiaohui Shen, and Liang Lin. Look into person: Joint body parsing & pose estimation network and a new benchmark. *PAMI*, 2018.

[19] Xingyu Liao, Lingxiao He, Zhouwang Yang, and Chi Zhang. Video-based person re-identification via 3d convolutional networks and non-local attention. In *ACCV*, 2018.

[20] Yutian Lin, Liang Zheng, Zhedong Zheng, Yu Wu, and Yi Yang. Improving person re-identification by attribute and identity learning. *arXiv:1703.07220*, 2017.

[21] Hao Liu, Jiashi Feng, Meibin Qi, Jianguo Jiang, and Shuicheng Yan. End-to-end comparative attention networks for person re-identification. *TIP*, 2017.

[22] Jinxian Liu, Bingbing Ni, Yichao Yan, Peng Zhou, Shuo Cheng, and Jianguo Hu. Pose transferrable person re-identification. In *CVPR*, 2018.

[23] Ting Liu, Tao Ruan, Zilong Huang, Yunchao Wei, Shikui Wei, Yao Zhao, and Thomas Huang. Devil in the details: Towards accurate single and multiple human parsing. *arXiv:1809.05996*, 2018.

[24] W Liu, X Chang, L Chen, and Y Yang. Semi-supervised bayesian attribute learning for person re-identification. In *AAAI*, 2017.

[25] Xihui Liu, Haiyu Zhao, Maoqing Tian, Lu Sheng, Jing Shao, Shuai Yi, Junjie Yan, and Xiaogang Wang. Hydraplus-net: Attentive deep features for pedestrian analysis. In *ICCV*, 2017.

[26] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *NIPSW*, 2017.

[27] Xuelin Qian, Yanwei Fu, Tao Xiang, Wenxuan Wang, Jie Qiu, Yang Wu, Yu-Gang Jiang, and Xiangyang Xue. Pose-normalized image generation for person re-identification. In *ECCV*, 2018.

[28] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *ECCV*, 2016.

[29] M. Saquib Sarfraz, Arne Schumann, Andreas Eberle, and Rainer Stiefelhagen. A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking. In *CVPR*, 2018.

[30] Arne Schumann and Rainer Stiefelhagen. Person re-identification by deep learning attribute-complementary information. In *CVPRW*, 2017.

[31] Yantao Shen, Hongsheng Li, Tong Xiao, Shuai Yi, Dapeng Chen, and Xiaogang Wang. Deep group-shuffling random walk for person re-identification. In *CVPR*, 2018.

[32] Yantao Shen, Hongsheng Li, Shuai Yi, Dapeng Chen, and Xiaogang Wang. Person re-identification with deep similarity-guided graph neural network. *ECCV*, 2018.

[33] Jianlou Si, Honggang Zhang, Chun-Guang Li, Jason Kuen, Xiangfei Kong, Alex C. Kot, and Gang Wang. Dual attention matching network for context-aware feature sequence based person re-identification. In *CVPR*, 2018.

[34] Chunfeng Song, Yan Huang, Wanli Ouyang, and Liang Wang. Mask-guided contrastive attention model for person re-identification. In *CVPR*, 2018.

[35] Chi Su, Jianing Li, Shiliang Zhang, Junliang Xing, Wen Gao, and Qi Tian. Pose-driven deep convolutional model for person re-identification. In *ICCV*, 2017.

[36] Chi Su, Fan Yang, Shiliang Zhang, Qi Tian, Larry Steven Davis, and Wen Gao. Multi-task learning with low rank attribute embedding for multi-camera person re-identification. *PAMI*, 2018.

[37] Chi Su, Shiliang Zhang, Junliang Xing, Wen Gao, and Qi Tian. Deep attributes driven multi-camera person re-identification. In *ECCV*, 2016.

[38] Arulkumar Subramaniam, Moitreya Chatterjee, and Anurag Mittal. Deep neural networks with inexact matching for person re-identification. In *NIPS*. 2016.

[39] Yumin Suh, Jingdong Wang, Siyu Tang, Tao Mei, and Kyoung Mu Lee. Part-aligned bilinear representations for person re-identification. In *ECCV*, 2018.

[40] Yifan Sun, Liang Zheng, Weijian Deng, and Shengjin Wang. Svdnet for pedestrian retrieval. In *ICCV*, 2017.

[41] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling. In *ECCV*, 2018.

[42] Chiat-Pin Tay, Sharmili Roy, and Kim-Hui Yap. Aanet: Attribute attention network for person re-identifications. In *CVPR*, 2019.

[43] Rahul Rama Varior, Bing Shuai, Jiwen Lu, Dong Xu, and Gang Wang. A siamese long short-term memory architecture for human re-identification. In *ECCV*, 2016.

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

[45] Cheng Wang, Qian Zhang, Chang Huang, Wenyu Liu, and Xinggang Wang. Mancs: A multi-task attentional network with curriculum sampling for person re-identification. In *ECCV*, 2018.

[46] Jingya Wang, Xiatian Zhu, Shaogang Gong, and Wei Li. Attribute recognition by joint recurrent learning of context and correlation. In *ICCV*, 2017.

[47] Jingya Wang, Xiatian Zhu, Shaogang Gong, and Wei Li. Transferable joint attribute-identity deep learning for unsupervised person re-identification. In *CVPR*, 2018.

[48] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, 2018.

[49] Yicheng Wang, Zhenzhong Chen, Feng Wu, and Gang Wang. Person re-identification with cascaded pairwise convolutions. In *CVPR*, 2018.

[50] Jing Xu, Rui Zhao, Feng Zhu, Huaming Wang, and Wanli Ouyang. Attention-aware compositional network for person re-identification. *CVPR*, 2018.

[51] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Deep metric learning for person re-identification. In *ICPR*, 2014.

[52] Rui Yu, Zhichao Zhou, Song Bai, and Xiang Bai. Divide and fuse: A re-ranking approach for person re-identification. *BMVC*, 2017.

[53] Yuan Yuhui and Wang Jingdong. Ocnet: Object context network for scene parsing. *arXiv:1809.00916*, 2018.

[54] 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 *CVPR*, 2017.

[55] Liming Zhao, Xi Li, Yueting Zhuang, and Jingdong Wang. Deeply-learned part-aligned representations for person re-identification. In *ICCV*, 2017.

[56] Rui Zhao, Wanli Ouyang, and Xiaogang Wang. Person re-identification by salience matching. In *ICCV*, 2013.

[57] Yiru Zhao, Xu Shen, Zhongming Jin, Hongtao Lu, and Xian-sheng Hua. Attribute-driven feature disentangling and temporal aggregation for video person re-identification. In *CVPR*, 2019.

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

[59] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. *arXiv:1701.07717*, 2017.

[60] Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. Re-ranking person re-identification with k-reciprocal encoding. In *CVPR*, 2017.

[61] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. *arXiv:1708.04896*, 2017.