
Usage Guidelines:
Please refer to usage guidelines at contact lib-eprints@bbk.ac.uk. or alternatively lib-eprints@bbk.ac.uk.
CDPM: Convolutional Deformable Part Models for Semantically Aligned Person Re-identification

Kan Wang, Changxing Ding, Stephen J. Maybank, and Dacheng Tao.

Abstract—Part-level representations are essential for robust person re-identification. However, common errors that arise during pedestrian detection frequently result in severe misalignment problems for body parts, which degrade the quality of part representations. Accordingly, to deal with this problem, we propose a novel model named Convolutional Deformable Part Models (CDPM). CDPM works by decoupling the complex part alignment procedure into two easier steps: first, a vertical alignment step detects each body part in the vertical direction, with the help of a multi-task learning model; second, a horizontal refinement step based on attention suppresses the background information around each detected body part. Since these two steps are performed orthogonally and sequentially, the difficulty of part alignment is significantly reduced. In the testing stage, CDPM is able to accurately align flexible body parts without any need for outside information. Extensive experimental results demonstrate the effectiveness of the proposed CDPM for part alignment. Most impressively, CDPM achieves state-of-the-art performance on three large-scale datasets: Market-1501, DukeMTMC-ReID, and CUHK03.

Index Terms—Person re-identification, alignment-robust recognition, part-based model, multi-task learning.

I. INTRODUCTION

PERSON re-identification (ReID) refers to the recognition of one pedestrian’s identity from images captured by different cameras. Given an image containing a target pedestrian (i.e., the query), a ReID system attempts to search a large set of pedestrian images (i.e., the gallery) for images that contain the same pedestrian. ReID has attracted substantial attention from both academia and the industry due to its wide-ranging potential applications, which include e.g. video surveillance and cross-camera tracking [1]. However, due to the large number of uncontrolled sources of variation, such as dramatic changes in pose and viewpoint, complex variations in illumination, and poor image quality, ReID remains a very challenging task.

The key to a robust ReID system lies in the quality of pedestrian representations. Many approaches [2], [3] attempt to directly extract holistic-level features from the whole image. These approaches typically suffer from overfitting problems [4]. Recently, part-level representations have been proven to be highly discriminative and capable of achieving state-of-the-art performance [5]–[9]. However, as illustrated in Fig. 1(a), the location of each body part varies in images due to errors in pedestrian detection [10], [11]. Consequently, the extracted part-level representations are not semantically aligned across images, meaning that they are not directly comparable for ReID purposes.

One intuitive strategy that could be adopted to resolve this issue involves directly detect body parts using additional tools, e.g., keypoints produced by pose estimation algorithms [12], [13], in both training and testing stages. However, the predictions made by these tools may not be sufficiently reliable, as they are usually trained on databases containing images that were captured under different conditions from images in ReID datasets. Another popular strategy involves detecting body parts via attention models that are seamlessly integrated into the ReID architecture [6], [14]–[16]. However, these attention models are optimized using the ReID task only; therefore, they are unable to provide explicit guidance for part alignment.

Accordingly, in this paper, we propose a novel framework for part alignment. By providing the training stage with a
minimal extra annotation (the upper and lower boundaries of the pedestrian) that is automatically detected, we are able to factorize the complicated part alignment problem into two simpler and sequential steps, i.e., a vertical alignment step that detects body parts in the vertical direction, and a horizontal refinement step which suppresses the background information around each detected part, as illustrated in Fig. 1(b).

Based on the above idea, we introduce a novel end-to-end model named Convolutional Deformable Part Models (CDPM), which can both detect flexible body parts and extract high-quality part-level representations. CDPM is built on a popular convolutional neural network (CNN) as backbone and further constructs three new modules, i.e., a feature learning module that extracts part-level features, a vertical alignment module that detects body parts in the vertical direction via multi-task learning, and a horizontal refinement module which is based on the attention mechanism.

Different channels in a CNN describe different visual patterns [4], [17]–[19], i.e., different body parts in the ReID context. In other words, channel-wise responses indicate hints of the location for each part. In light of the above, CDPM succinctly integrates these three modules, based on the output of the same backbone model. In the inference stage, the vertical alignment module and the horizontal refinement module run sequentially for part alignment, followed by high-quality feature extraction from the aligned parts.

The effectiveness of the proposed method is systematically evaluated on three popular ReID databases, i.e., Market-1501 [20], DukeMTMC-reID [21], and CUHK03 [22]. Experimental results demonstrate that CDPM consistently achieves superior performance and outperforms other state-of-the-art approaches by a considerable margin.

We summarize the contributions of this work as follows:

- We formulate the novel idea of decoupling the body-part alignment problem into two orthogonal and sequential steps, i.e., a vertical detection step and a horizontal refinement step. These two steps establish a novel framework for the learning of high-quality part-level representations. To the best of our knowledge, this is the first attempt to solve the misalignment problem by means of decomposition into orthogonal directions.
- Under the divide-and-conquer formulation, we propose a succinct CDPM architecture that integrates representation learning and part alignment through sharing of the same backbone model. In particular, the vertical alignment module is realized by an elaborately designed multi-task learning structure.
- Extensive evaluation on three large-scale datasets demonstrate the superiority of the proposed CDPM. We further conduct comprehensive ablation study to enable analysis of the effectiveness of each component of CDPM.

The remainder of this paper is organized as follows. We first review the related works in Section II. Then, we describe the proposed CDPM in more detail in Section III. Extensive experimental results on three benchmarks are reported and analyzed in Section IV, after which the conclusions of the present work are outlined in Section V.

A. Person Re-identification

Prior to the prevalence of deep learning, approaches to ReID could be divided into two distinct categories, namely, feature engineering methods [23], [24] and metric learning methods [25]–[29]. Over the past few years, deep learning-based approaches [4]–[9], [12], [13], [30]–[40] have come to dominate in the ReID community. Many works target the learning of discriminative representations from the holistic image directly. One common strategy involves training deep models to learn ID-discriminative embedding (IDE) as an image classification task [2]. Moreover, the quality of image representations can be enhanced using metric learning-based loss functions, such as contrastive loss [41], triplet loss [42], and quadruplet loss [43]. Other works train deep models using a combination of different types of loss functions [44], [45].

The holistic image-based approaches typically suffer from overfitting problems [4]. To relieve this issue, part-based approaches [4]–[8], [14] have been proposed in order to learn discriminative image representations. However, due to errors that commonly arise in pedestrian detection context, the location of each body part varies in different images. When considering strategies for handling this part misalignment problem, existing approaches can be grouped into three categories.

1) Pre-defined Part Location-based Methods: These approaches extract part-level features from patches [46] or horizontal stripes [5], [8], [47] of pre-defined locations. For example, Cheng et al. [47] uniformly divide a pedestrian image into four horizontal stripes, from which part-level features are then extracted. Wang et al. [5] also partition one image into horizontal stripes. It mitigates the part misalignment problem by extracting multi-granularity part-level features. However, as the above methods usually assume that the misalignment problem is moderate; they may therefore encounter difficulties when handling cases of severe misalignment.

2) Outside Information-based Methods: These methods align body parts through the use of outside information, e.g., masks produced by human parsing tools [48]–[50] or key points detected by pose estimation algorithms [12], [13], [51]. In these approaches, outside information is usually required in both the training and testing stages [12], [13], [50]. The downsides to these methods are first, there is additional computational cost, second, the accuracy of part alignment depends on the performance of the outside tools, which are usually trained on databases made up of images captured under conditions that are significantly different from ReID databases.

3) Attention Model-based Methods: These methods learn to directly predict bounding boxes or soft masks for body parts from feature maps produced by ReID networks, without making use of any additional supervision [6], [14], [16], [17]. For example, Li et al. [6] designed a hard regional attention model that is able to predict bounding boxes for each body part. In comparison, Zhao et al. [16] proposed to predict a set of soft masks. Element-wise multiplication between one soft mask and each channel of feature maps is used to produce part-level features. However, the lack of explicit supervision of
part alignment may cause difficulties during the optimization of these attention models.

The proposed CDPM approach improves the accuracy of part alignment by introducing a minimal amount of extra supervision in the training stage, through which the complicated part alignment problem can be decomposed into two separate and simpler steps. Therefore, compared with the attention-based methods, the optimization difficulty associated with part alignment is significantly reduced. Compared with the second category of methods, the utilized annotations are more robust. Besides, CDPM does not require any outside information in the testing stage; therefore it is easier to use in practice.

### B. Part-based Object Detection

Prior to the emergence of deep learning as a widespread phenomenon, the Deformable Part Model (DPM) was one of the most popular methods for object detection. In both DPM [52] and its deep versions [53]–[55], part detection is performed as an auxiliary task to promote detection accuracy. Over the past few years, region proposal-based methods [56], [57] have become more popular. Unlike DPM methods, region proposal-based methods usually detect the whole object directly rather than engaging in explicit part detection.

By contrast, the proposed method aims to detect flexible parts only, as the coarse location of the whole body is already known. From this perspective, our method bears more similarity to DPM than to region proposal-based methods. Since our proposed method is based on CNN, we name it Convolutional Deformable Part Models (CDPM).

### III. CONVOLUTIONAL DEFORMABLE PART MODELS

#### A. Problem Formulation

As illustrated in Fig. 1, we decouple the complex part alignment problem into two separate and sequential steps, i.e., a vertical alignment step that locates each body part in the vertical direction, and a horizontal refinement step which suppresses the background information around each body part.

The first step is more challenging. This is because, firstly, the whole image is searched for each body part; secondly, there is usually no clear boundary between adjacent parts. We meet the above challenges by providing a minimal extra annotation as a form of auxiliary supervision in the training stage.

As shown in Fig. 2(a), the upper and lower boundaries of pedestrians are automatically detected by Macro-Micro Adversarial Network (MMAN) [58]. The upper and lower boundaries of pedestrians are obtained according to the following rules. First, the seven classes obtained by MMAN [58] are merged into three categories, i.e., the head (including the hair and face), the upper body (consisting of the upper-clothes and arms), and the lower part of the body (comprises lower-clothes, legs and shoes). Second, the upper boundary of the pedestrian is set as the upper boundary of the head, while the lower boundary of this pedestrian is defined as the lower boundary of the lower part of the body. Besides, as shown in Fig. 2(b), severe part-missing problem is detected by counting the number of pixels for the head and the lower-part of the body, respectively. If the size of either of them is smaller than a pre-set threshold (e.g., 1280 pixels in our implementation), we regard this image as having a severe part-missing problem. Since MMAN may fail to produce reliable results (Fig. 2(c)), it is suboptimal to directly utilize the location of body parts returned by MMAN.

The equal partition between these two detected boundaries produces the area of each part in the vertical direction. It is worth noting here that the annotation is not required in testing; therefore, the detected annotation can be regarded as a kind of privileged information [59]. Moreover, the second step is comparatively much easier; therefore, it is not provided with any extra information.

Based on the above ideas, we propose our novel CDPM model for joint part feature learning and part alignment. As illustrated in Fig. 3, CDPM is built on the ResNet-50 backbone model [60]. Similar to [7], we remove the last spatial downsampling operation in ResNet-50 so as to increase the size of the output feature maps. Based on these output feature maps, we go on to construct three new modules, i.e., the feature learning module for part-level feature extraction, the vertical alignment module based on multi-task learning, and the horizontal refinement module based on attention. These three modules work collaboratively together to align body parts and further learn high-quality part-level representations.

#### B. Feature Learning Module

The feature learning module is based on one recent work named Part-based Convolutional Baseline (PCB) [7]. As illustrated in Fig. 3, the feature learning module incorporates \( K \) part-level feature learning branches, each of which learns part-specific features. These \( K \) branches all have an identical structure, i.e., one Global Average Pooling (GAP) layer, one \( 1 \times 1 \) convolutional layer, and one classification layer. In the training stage, the location of each part can be inferred via the provided annotations pertaining to the upper and lower pedestrian boundaries by means of uniform partition (Fig. 2(a)). If the upper or lower boundary is not provided, e.g., in cases where they are invisible due to part-missing problems.

![Fig. 2. (a) The automatically detected upper and lower boundaries for a pedestrian, which are denoted as the green line and the blue line, respectively. The location of each part is inferred through the uniform partition between the two boundaries. (b) Severe part-missing problem is detected by counting the number of pixels for the head and the lower-part of the body, respectively. Best viewed in color. (c) The parsing results may not be sufficiently reliable for low-quality images.](image-url)
C. Vertical Alignment Module

Different channels in the feature maps produced by the ReID network describe different body parts [4], [18]. This indicates that the channel-wise responses can provide hints as to part location. We therefore design a detection module to locate body parts in the vertical direction, based on the output of the backbone model only. In the interests of simplicity, we divide the output of the backbone model into \( R \) sliding windows, each with fixed height and width. We then select one sliding window for each part using the proposed vertical alignment module. In this paper, the size of output of the backbone model is \( 24 \times 8 \times 2048 \), where these three dimensions denote height, width, and channel number, respectively. The size of each sliding window is set as \( 4 \times 8 \times 2048 \); therefore \( R \) is equal to 21 for each image (Fig. 4(b)).

Furthermore, inspired by Faster R-CNN [57], we process the sliding windows by means of multi-task learning, i.e., coarse classification of all sliding windows, and the refined regression of sliding windows to their respective ground-truth locations. These two tasks share only one GAP layer. It is worth noting that we only utilize images whose upper and lower boundaries are both visible during the training for this module.

1) Coarse Classification of Sliding Windows: This task classifies all sliding windows to their corresponding parts or the background category. To do so, it incorporates one \( 1 \times 1 \) convolutional layer and one fully connected (FC) layer. The loss function for the feature learning module is as follows:

\[
\mathcal{L}_f = \sum_{k=1}^{K} \mathcal{L}_p^k, \tag{2}
\]

where \( \mathcal{L}_p^k \) is the loss for the \( k \)-th part classification.

Fig. 3. Architecture of the proposed CDPM. Based on the ResNet-50 backbone model, CDPM constructs three new modules, i.e., the feature learning module including \( K \) part-level branches, the vertical alignment module, and the horizontal refinement module. In the inference stage, the vertical alignment module classifies all sliding windows to their corresponding parts or the background category. To do so, it incorporates one \( 1 \times 1 \) convolutional layer and one fully connected (FC) layer. The proposed part alignment method, thereby enabling the part-level feature learning branches to learn robust features.
important parameters for the two layers are outlined in Fig. 3. The output of the FC layer is \((K+1)\)-dimensional, which denotes \(K\) parts and the background category. As each of these sliding windows may overlap with two adjacent parts, their ground-truth labels are soft rather than one-hot. We therefore optimize the classification task using the sigmoid cross-entropy loss, which can be formulated as \(L_c = \)

\[
- \frac{1}{NR} \sum_{i=1}^{N} \sum_{k=0}^{R} \sum_{r=1}^{K+1} y_i^{r(k)} \log \hat{y}_i^{r(k)} + (1 - y_i^{r(k)}) \log(1 - \hat{y}_i^{r(k)})],
\]

where \(y_i^{r(k)}\) is the ground-truth probability of the \(r\)-th sliding window belonging to the \(k\)-th part of the \(i\)-th image, and \(\hat{y}_i^{r(k)}\) signifies the corresponding predicted probability value.

**Ground-truth Label of Sliding Windows** For a given sliding window \(r(u_r, l_r)\) with upper and lower boundaries \(u_r\) and \(l_r\), respectively, we associate this window with a ground-truth label vector \(y^r = (y^{r(1)}, y^{r(2)}, ..., y^{r(K)}, y^{r(K+1)})\). The value of each element in \(y^r\) depends on the size of the overlap between \(r\) and the corresponding body part or background category. In more detail, we first calculate the ground-truth upper and lower boundaries for the \(k\)-th part:

\[
u_k = U + (k - 1) \times \frac{V - U}{K},
\]

\[
l_k = u_k + \frac{V - U}{K},
\]

where \(U\) and \(V\) represent the annotated upper and lower boundaries of the pedestrian in the training image. The ground-truth area for the \(k\)-th part is denoted as \(p_k(u_k, l_k)\). Then,

\[
y^{r(k)} = \frac{S(p_k(u_k, l_k) \cap r(u_r, l_r))}{S(r(u_r, l_r))}, 1 \leq k \leq K,
\]

and \(y^{r(K+1)} = 1 - \sum_{k=1}^{K} y^{r(k)}\), where \(S(*)\) denotes the size of the area \(*\).

2) **Refined Regression of Sliding Windows:** We further promote the accuracy of the vertical alignment module by means of part-specific regression tasks. As illustrated in Fig. 2, the \(K\) regression tasks are constructed so that they all have the same structure. However, these tasks do not share any parameters, and each task is optimized for the detection of one specific part.

Each regression task incorporates one channel attention block [61], one \(1 \times 1\) convolutional layer, one FC layer, and one tahn layer. The important parameters for the above layers are labeled in Fig. 5. The channel attention block is used to highlight the information for one specific part. Each channel attention block includes two successive \(1 \times 1\) convolutional operations, whose important parameters are labeled in Fig. 5. The output of the sigmoid layer is channel attention. The input feature vector for the attention block is then multiplied with channel attention in an element-wise manner. Finally, we obtain the weighted feature vector for each sliding window.

During training, the 2048-dimensional features of all sliding windows are fed into the \(K\) regression branches. Correspondingly, we obtain \(K\) sets of predicted offsets after tanh normalization. Each regression task involves the optimization of the Mean Squared Error (MSE) loss:

\[
\mathcal{L}_r^k = \frac{1}{2R^k} \sum_{i=1}^{N} \sum_{j=1}^{R} (\Delta_j^{i(k)} - \hat{\Delta}_j^{i(k)})^2 \cdot 1\{|\Delta_j^{i(k)}| < 1\},
\]

where \(\Delta_j^{i(k)}\) denotes the ground-truth offset of the \(j\)-th sliding window for the \(k\)-th part in the \(i\)-th image, while \(\hat{\Delta}_j^{i(k)}\) is the corresponding predicted value. \(1\{|\Delta_j^{i(k)}| < 1\}\) is equal to either 0 or 1, meaning that we only utilize sliding windows with a ground-truth offset value within the range of \((-1, 1)\). \(R^k\) denotes the number of sliding windows that satisfy \(1\{|\Delta_j^{i(k)}| < 1\}\) for all \(k\)-th parts in a mini-batch. \(\Delta_j^{i(k)}\) can be easily obtained by first subtracting the sliding window’s center coordinate in the vertical direction from that of the \(k\)-th part (Eq. 5), and then normalized by the height of the sliding window. Finally, the joint loss function for the vertical alignment module can be formulated as follows:

\[
\mathcal{L}_v = \mathcal{L}_c + \sum_{k=1}^{K} \mathcal{L}_r^k.
\]

During testing, the 2048-dimensional features of all sliding windows are fed into the classification task and \(K\) regression tasks simultaneously. The classification scores and predicted offsets for the \(R\) sliding windows are obtained for each part. We select the optimal sliding window for each part according to the following rules: first, if the classification scores of multiple sliding windows are above a pre-defined threshold \(T\), we select the one with the smallest offset (absolute value); second, if there is only one or no sliding window/s with a classification score above \(T\), we simply choose the one with the largest classification score.

**D. Horizontal Refinement Module**

It must be reiterated here that the above module only detects body parts in the vertical direction. Accordingly, we further propose to suppress the background information in the \(K\) selected sliding windows via an additional horizontal refinement operation. As shown in Fig. 2, the horizontal refinement module is applied to each part-level feature learning branch. In this paper, we realize this module using the Spatial-Channel Attention (SCA) model proposed in [6].

For completeness sake, we briefly introduce the structure of SCA. As shown in Fig. 6, the spatial and channel attentions
of SCA are realized by separate branches: the former branch comprises a global cross-channel average pooling layer, a convolutional layer, a resizing bilinear layer, and another convolutional layer; the latter branch consists of a GAP layer and two successive convolutional layers. Finally, these two types of attention information are fused by one convolutional layer and normalized by one sigmoid layer. The important parameters of SCA are marked in Fig. 6, while additional details regarding implementation can be found in [6].

It is worth noting here that SCA was originally employed to suppress the background information in the holistic pedestrian image [6], rather than around each body part. We would argue that our \textit{divide-and-conquer} strategy is more intuitive and effective, since it makes it dramatically easier to distinguish pixels of a single part from the surrounding background pixels. By comparison, applying SCA to a holistic image can be much more difficult, as both the structure of the whole body and the background information in the holistic image are often significantly more complicated.

E. Person Re-ID via CDPM

In the training stage, and taking all three modules of CDPM into account, the overall objective function for CDPM can be written as follows:

\[
\mathcal{L} = \mathcal{L}_f + \mathcal{L}_v
\]

where \(\lambda_1\) and \(\lambda_2\) are the weights of the loss functions. For simplicity, these are consistently set to 1 in this paper.

In the testing stage, each image passes through the backbone model to yield feature maps of size \(24 \times 8 \times 2048\). For extracting the part-level features, these \(24 \times 8 \times 2048\) feature maps are divided into \(R\) sliding windows, the size of which is fixed to \(4 \times 8 \times 2048\). The vertical alignment module selects one optimal sliding window for each part; subsequently, the selected sliding window for the \(k\)-th part passes through the \(k\)-th horizontal refinement module and part-level feature learning branch in order to obtain the 512-dimensional part-level feature vector \(z^k\). The final representation of the image is obtained by concatenating the above \(K\) feature vectors:

\[
f = [z^1, z^2, ..., z^K].
\]
CUHK03 includes 14,097 images of 1,467 identities. Images of each identity were captured by two disjoint cameras. This dataset provides both hand-labeled and DPM-detected bounding boxes. We evaluate our method using both types of bounding boxes. Besides, we adopt the new train/test protocol proposed in [62]. The new protocol splits the dataset into a training set of 767 identities and a testing set of the remaining 700 identities.

B. Implementation Details

The naive combination of the backbone model and the feature learning module, i.e., the PCB model [7], is selected as the baseline. Compared with CDPM, the baseline model lacks the part alignment ability. We here uniformly divide the feature maps produced by the backbone model in the vertical direction to create K non-overlapped parts, which become the input for the part-level feature learning branches.

1) Hyper-parameters of CDPM: The number of body parts, i.e., K, is set to 6 following the baseline model [7]. Besides, the threshold value T for sliding window selection is empirically set to 0.60 for Market-1501 and 0.35 for the other two databases. The value of the hyper-parameters of SCA are kept the same as in the original work [6]. Finally, when triplet loss is utilized for training, we set P to 6 and A to 8, while the margin α for the triplet loss is set to 0.4.

2) Training details: Experiments are conducted using the PyTorch framework. All pedestrian images are resized to 384 × 128 pixels. In line with existing works [22], [63], we adopt extensive data augmentation to reduce overfitting. Specifically, we augment the training data via offline translation [22], online horizontal flipping, and online random erasing [63]. After the offline translation, the size of each training set is enlarged by 5 times. Moreover, the ratio of random erasing is set to 0.5.

We use the standard stochastic gradient descent with momentum [64] for model optimization and further set the momentum value to 0.9 and the batch size N to 48. Besides, we utilize a stage-wise strategy to train the proposed CDPM. In the first stage, we fine-tune the baseline model from the IDE model proposed in [2] for 50 epochs; here, the learning rate is initially set to 0.01 and then multiplied by 0.1 for every 20 epochs. In the second stage, we fix the parameters of the baseline model and optimize only the parameters of the components newly introduced in CDPM (i.e., the vertical alignment module and horizontal refinement modules). This stage is trained for 40 epochs, with the learning rate set to 0.01 initially and then multiplied by 0.1 for every 15 epochs. Finally, all CDPM model parameters are fine-tuned in an end-to-end fashion for 30 epochs, with a small initial learning rate of 0.001 that decreases to 0.0001 after 20 epochs.

C. Comparisons to State-of-the-Art Methods

The essential contribution of CDPM lies in its ability to detect flexible body parts for ReID. For fair comparison with existing approaches, we categorize them into three groups, i.e., holistic feature-based methods, single-level part feature-based methods, and multi-granularity feature-based methods. Hereafter, they are denoted as ‘HF-based’, ‘SPF-based’, and ‘MGF’, respectively.
1) Evaluation on Market-1501 Dataset: Performance comparisons between CDPM and state-of-the-art methods on the Market-1501 database are tabulated in Table I. From the table, it can be seen that CDPM significantly outperforms all existing methods for both Rank-1 and mAP results, and achieves state-of-the-art performance. In particular, CDPM outperforms the most recent SPF-based method, i.e., PCB+RPP [7], by 1.4% and 4.4% on the single-query mode for Rank-1 accuracy and mAP, respectively. The above comparisons successfully demonstrate the effectiveness of CDPM. Moreover, the performance of CDPM is further improved through the extraction of multi-granularity features. Finally, CDPM* achieves state-of-the-art performance superior to that of all other approaches. Specifically, CDPM* achieves 95.9% and 87.2% for Rank-1 accuracy and mAP on the single-query mode, respectively.

2) Evaluation on DukeMTMC-ReID Dataset: Compared with Market-1501, the pedestrian images in the DukeMTMC-ReID database are impacted by more variations in viewpoint and background. Performance comparisons are summarized in Table II. These results reveal that CDPM achieves the best Rank-1 accuracy and mAP overall, outperforming all other state-of-the-art methods by a large margin. In particular, CDPM outperforms the best existing SPF-based approach [68] by 3.8% and 8.2% for Rank-1 accuracy and mAP, respectively. These results suggest that CDPM can locate body parts accurately, even in cases where dramatic variations exist in terms of the viewpoints and background.

3) Evaluation on CUHK03 Dataset: We evaluate the performance of CDPM on CUHK03 using both manually labeled and auto-detected bounding boxes. The results of the comparison are tabulated in Table III. From the table, it can be seen that CDPM achieves the best performance among all SPF-based approaches. In particular, it outperforms the second-best approach [7] by 8.2% and 9.5% for Rank-1 accuracy and mAP using auto-detected bounding boxes, respectively. Furthermore, CDPM* also achieves the best performance among MGF-based approaches. These above comparisons clearly justify the overall effectiveness of CDPM.

D. Ablation Study

In the following, we present the results of ablation study conducted to justify the effectiveness of each newly introduced component of CDPM, i.e., the vertical alignment module and horizontal refinement module. Moreover, we also compare some possible variants of the vertical alignment module and design several experiments to evaluate the influence of the annotation on the performance of CDPM. At last, efficiency comparison and some visualizations are performed to further verify the effectiveness and superiority of CDPM. In line with recent works [6], [7], the ablation study is conducted on both the Market-1501 and DukeMTMC-ReID datasets.

1) Effectiveness of the Vertical Alignment Module: In this subsection, we equip the baseline model with the vertical alignment module and denote this model as Baseline+V in Table IV. As shown in Table IV, equipping the vertical alignment module consistently promotes ReID performance. In particular, Baseline+V outperforms the baseline model by 1.1% and 1.5% in terms of Rank-1 accuracy on Market-1501 and DukeMTMC-ReID, respectively. Moreover, exemplars of body-part detection results achieved by the vertical alignment module are presented in Fig. 9. It can be clearly seen from the
In this subsection of the experiments, we equip the baseline detection scheme of CDPM. Clearly, demonstrate the superiority of the proposed part-based parts is independent in Baseline+V. Therefore, Baseline+V of all body parts. In comparison, the detection of either of the two boundaries will cause errors in the location is less reliable in body part detection. Detection errors in 2.3% in mAP, respectively. The main reason is Baseline+V(G) surpassing Baseline+V(G) by 0.9% in Rank-1 accuracy and 84.5% in mAP on the Market-1501 database, line+V(G). For example, Baseline+V obtains 94.6% in Rank-1 accuracy on Market-1501 from 93.5% to 94.7%, which represents a 18.5% relative reduction in the error rate.

Moreover, similar to [6], we also try to apply only one SCA module to the whole feature maps output by the backbone model; this approach is denoted as Baseline+H(G) in Table IV. Experimental results indicate that Baseline+H consistently outperforms Baseline+H(G). This result justifies our motivation that it is much easier to reduce the interference of background information in a divide-and-conquer manner.

Finally, we simultaneously equip the baseline model with both modules. The results can be found in the row ‘CDPM’ in Table IV. From these results, it is clear that the combination of the two modules creates a considerable performance boost relative to the use of one module. Compared with the baseline model, CDPM boosts the Rank-1 accuracy by 1.7% and 2.2%, and mAP by 4.1% and 3.0%, on the Market-1501 and DukeMTMC-ReID datasets, respectively. From the above comparisons, we conclude that the vertical alignment module and horizontal refinement module are complementary; therefore the proposed divide-and-conquer solution is effective.

3) Structure of the Vertical Alignment Module: In this subsection, we compare the performance of the vertical alignment module with two possible variants. The first variant (denoted as Variant 1 in Table V), which is similar to Faster-RCNN [57], shares the parameters of the 1 × 1 convolutional layer of the coarse classification task and K refined regression tasks. The second variant (denoted as Variant 2 in Table V) shares the parameters of the 1 × 1 convolutional layer between each part-level feature learning branch and the corresponding refined regression task in the vertical alignment module. As can be seen from Table V, both of these variants are inferior to the proposed CDPM. For example, the Rank-1 accuracies of the two variants on the Market-1501 database are lower than those achieved by CDPM by 0.9% and 3.0%, respectively.

Two insights can be gained from the above results: firstly, in contrast with object detection [57], body part detection is a fine-grained task, meaning that more independent parameters

---

**Fig. 9.** The vertical alignment module can detect body parts robustly under a vast majority of circumstances, including moderate (a, b) or even severe (c, d) part misalignment, dramatic pose variations (e, f), and occlusion (g). However, the vertical alignment module may fail in cases where complex occlusion or a part missing problem exist (h). Best viewed in color.
are required for each specific part detection task; secondly, part-level feature learning for recognition and body part detection are heterogenous tasks, in that the former learns the unique characteristic of one identity while the latter is based on the general characteristic of one specific body part between different identities. Therefore, these two tasks should not share parameters.

4) Comparisons in Different Types of Annotations: In the following, we explore the influence of the quality of annotations by comparing the performance of the proposed CDPM when trained with automatic annotations as opposed to manual annotations.

From Fig. 10, it is clear that the performance of CDPM on both the Market-1501 and DukeMTMC-ReID databases are fairly similar regardless of which type of annotations is used. For example, compared with manual annotations, the performance of CDPM with automatic annotations is reduced by only 0.1% in Rank-1 accuracy and 0.2% in mAP on the Market-1501 database. These results demonstrate that CDPM is robust to the quality of annotations, meaning that it will be easy to use in real-world applications.

5) Efficiency Comparison: We compare the efficiency of CDPM with some state-of-the-art works [5], [7], [12], [13], then tabulate the results of this comparison in Table VI. It is worth noting here that the existing works are quite different as regards the details of their implementation. Therefore, for fair comparison, we mainly draw comparisons with approaches that adopt similar backbone models. We also resize the input image for all models in Table VI so that the image size is as close to 384 × 128 pixels as possible.

It is also worth noting here that neither PCB [7] or MGN [5] require part detection; they are therefore faster. In comparison, Spindle [12], PDC [13], and CDPM perform part detection and part-level feature extraction separately; therefore, they require more computational costs. As shown in Table VI, CDPM is competitive in terms of efficiency when compared with Spindle [12] and PDC [13]. For example, the time cost of CDPM in the testing stage is only about 56.7% of PDC’s [13]. This is because CDPM integrates part detection and part-level feature extraction into one compact framework.

6) Visualization of Features for the Vertical Alignment Module: To explain the effectiveness of the vertical alignment module, we visualize the features learned by vertical alignment module utilizing t-SNE [71]. These features are extracted from the last convolutional layer of each of the K regression tasks in the vertical alignment module for each body part. We select five images in each of 50 identities from the gallery set of the Market-1501 database. These selected images are affected to greater or lesser extents by the part misalignment problem.

From Fig. 11, we can also clearly see that the features from the same body part are close to each other, despite being from different identities. Moreover, we also observe that the inter-part difference is significant when compared with intra-part variances. These results verify that the vertical alignment module is able to capture the general characteristic of each body part.

7) Visualization of Part-relevant Saliency Maps: To arrive at an interpretation of how the proposed CDPM captures the body parts, we opt to visualize the part-relevant saliency maps generated by each feature learning module of CDPM utilizing Grad-CAM [72].

The resulting saliency maps are presented in Fig. 12. We
can also clearly observe from the figure that CDPM adaptively focuses on discriminative body parts, even when these the body parts are misaligned due to detection error (Fig. 12(a)(b)) or variation of pose (Fig. 12(c)). These visualization results verify the effectiveness of the proposed CDPM in solving the part misalignment problem.

8) Visualization of Retrieval Results: Finally, we visualize the retrieval results of CDPM and two other state-of-the-art approaches, namely PCB [7] and MGN [5].

As illustrated in Fig. 13, CDPM is able to yield more reliable results than both PCB and MGN for images with part misalignment problems (Fig. 13(a)(b)(d)(e)); this is due to its ability to flexibly align the body parts between images. Moreover, we can also observe that CDPM is more robust to both occlusion (Fig. 13(c)) and part-missing problems (Fig. 13(f)) than the other two approaches. These results further demonstrate the effectiveness of the proposed CDPM.

V. CONCLUSION

In this work, we study the part misalignment problem in ReID and propose a novel framework named CDPM, which integrates part-level feature learning and part alignment in a succinct model. In a departure from existing works, we decouple the complicated part misalignment problem into two orthogonal and sequential steps: the first step detects body parts in the vertical direction, while the second step separately refines the boundary of each body part in the horizontal direction. Thanks to the divide-and-conquer strategy, each of these two steps becomes significantly simpler. We conduct extensive experiments on three large-scale ReID benchmarks, through which the effectiveness of the proposed model is comprehensively justified and state-of-the-art performance is achieved. We also conduct detailed ablation study to prove the effectiveness of each component in the proposed model.

REFERENCES

Fig. 13. Examples of the retrieval results on the Market-1501 database. In each group of images, the leftmost one represents the query image, while the remaining images in the first, second, and third rows are the top-5 retrieval results from CDPM, PCB [7], and MGN [5], respectively. Green rectangles denote true positives, while red rectangles indicate the false positives. Best viewed in color.


