Learning Disentangled Representation Implicitly Via Transformer for Occluded Person Re-Identification

Mengxi Jia, Xinhua Cheng, Shijian Lu, and Jian Zhang, Member, IEEE

Abstract—Person re-Identification (re-ID) under various occlusions has been a long-standing challenge as person images with different types of occlusions often suffer from misalignment in image matching and ranking. Most existing methods tackle this challenge by aligning spatial features of body parts according to external semantic cues or feature similarities but this alignment approach is complicated and sensitive to noises. We design DRL-Net, a disentangled representation learning network that handles occluded re-ID without requiring strict person image alignment or any additional supervision. Leveraging transformer architectures, DRL-Net achieves alignment-free re-ID via global reasoning of local features of occluded person images. It measures image similarity by automatically disentangling the representation of undefined semantic components, e.g., human body parts or obstacles, under the guidance of semantic preference object queries in the transformer. In addition, we design a decorrelation constraint in the transformer decoder and impose it over object queries for better focus on different semantic components. To better eliminate interference from occlusions, we design a contrast feature learning technique (CFL) for better separation of occlusion features and discriminative ID features. Extensive experiments over occluded and holistic re-ID benchmarks show that the DRL-Net achieves superior re-ID performance consistently and outperforms the state-of-the-art by large margins for occluded re-ID dataset.

Index Terms—Occlusion scene, person Re-identification, representation learning, visual Transformer.

I. INTRODUCTION

PERSON re-IDentification (re-ID) [1] is a computer vision task that aims to associate person images captured by non-overlapping cameras. It has been studied intensively in recent years due to its wide applications in various video surveillance tasks [2]–[6]. Thanks to the advance in deep learning and large-scale benchmarks, the re-ID research has achieved substantial progress and different approaches have been successfully proposed to tackle variations in viewpoints and poses [7], illumination conditions [8], camera configurations [9], etc. On the other hand, most existing holistic re-ID methods [10] assume that the entire human body is visible in person images which hence cannot generalize well to occluded person images that suffer from incomplete information with various invisible body parts. Since humans are often occluded by clutters and obstacles in natural scenes, occluded re-ID [11], [12] has great values in different surveillance tasks which is worth further investigation despite the complication resulting from missing body-part information.

Occluded re-ID is facing two major challenges. The first is super-rich variation of occlusions that block different body parts randomly and change the appearance of person images substantially. The occlusions thus introduce more intra-class variations which lead to more image matching errors and degraded re-ID performance. The second is interference of occlusions which often shares similar appearance as body parts and deteriorates the learnt person image representations. Most existing methods address the occlusion challenge by detecting the non-occluded body parts and aligning visible human parts in person image matching, and two typical alignment approaches have been widely investigated. The first approach exploits various external cues such as person masks [13], [14], semantic parsing [15] and pose estimation [12], [16], [17] for accurate alignment of visible body parts. However, the extraction of external cues is sensitive which tends to fail while facing severe occlusions and background noises. The other approach aligns body parts based on the similarity of local image features [18]–[21], but it often struggles in differentiating human bodies from obstacles which often leads to mismatches. Beyond that, both alignment approaches involve complicated extra operations that take time in inference and also tend to accumulate errors.

This paper presents DRL-Net, an alignment-free re-ID framework that handles occlusions through disentangled representation learning as illustrated in Fig. 1. Leveraging the transformer architecture [22], DRL-Net eliminates the error-prone person alignment operations which first extracts compact image representations using CNNs and then performs global reasoning and ID prediction using the transformer encoder and decoder. Specifically, DRL-Net disentangles the representations of undefined semantic components in occluded person images based on object queries without any additional supervision. Under the guidance of semantic preferences object queries, it adapts
We propose a novel transformer framework DRL-Net that tackles occluded person re-ID by learning disentangled representation implicitly without any additional supervision and complicated alignment process. For the transformer decoder, we impose a decorrelation constraint over semantic preference object queries to force them to focus on respective semantic components. In addition, we design a contrast feature learning module and a data augmentation strategy for better isolating ID-irrelevant features from global representation and suppressing occlusion interference.

The main contributions of this work are three-fold.

- We propose a novel transformer framework DRL-Net that tackles occluded person re-ID by learning disentangled representation implicitly without any additional supervision and complicated alignment process.
- We design a novel contrast feature learning technique together with a data augmentation strategy that mitigate the interference of occlusion noises effectively.
- The proposed DRL-Net achieves state-of-the-art performance under various occlusions yet without sacrificing performance over normal re-ID data with little occlusion.

II. RELATED WORK

Person Re-ID has been one of the most studied problems due to its important application, and most of existing works were developed for matching holistic person that cannot tackle the occluded Re-ID problem. Since our method is proposed for occluded re-ID and based on transformer architecture, we only briefly review several related works in this section.

A. Occluded Person Re-ID

The challenges of occluded re-ID mainly lie in body information incompleteness and spatial misalignment. Existing occluded re-ID methods can be roughly summarized into two streams, approaches with external cues and approaches based on part-to-part matching.

Previous methods leverage external cues such as human parsing, pose estimation or foreground segmentation to align parts of bodies. Under the guidance of extra semantic labels, such methods align parts precisely and benefit the feature representation. Miao et al. [12] propose a pose-guided feature alignment method (PGFA), taking advantage of the human semantic key-points to guide the matching of probe and gallery images. Gao et al. [16] present a pose-guided visible part matching algorithm (PVPM) which jointly learns features and predicts the part visibility with attention heatmaps guided by pose estimation and graph matching accordingly. Wang et al. [17] propose a framework jointly modeling high-order relation and human-topology information by utilizing key-points estimation for robustly aligned features. However external cues requiring limits their usage and robustness in practical deployment. The inference of extra modules costs more time inevitably, and the generated semantic labels are untrustworthy under severe occlusions or low-resolution scenarios.

Models based on Part-to-part matching strategy handle occlusions by generating part alignment relations according to the similarity of local features across query and gallery images. Sun et al. [23] propose a network named Part-based Convolutional Baseline (PCB) which divided feature maps into horizontal pieces to learn local features directly. Zhang et al. [24] align local features and compute distance by finding the shortest path. Zhu et al. [25] locate human body parts and potential person belongings at pixel-level by clustering algorithms to alignment. These methods match local features through self-supervision without external cues. Nevertheless, such auto-alignment steps require complicated algorithms like shortest path finding and clustering, and predict results strongly influenced by the way of dividing images. Different from above strict alignment-based approaches, our method addresses the occluded person re-ID by leveraging transformer architectures, which can automatically and implicitly extract and disentangle the representations of target person without any additional supervision.

B. Visual Transformer

Transformer is a type of deep neural network which utilizes the self-attention mechanism and shows great performance on natural language processing tasks. Inspired by the significant success of transformer in the NLP field [26]–[29], researchers applied transformer to various computer vision areas. Carion et al. [30] presented detection transformer (DETR) to view object detection as a direct set prediction problem, which firstly bring transformer architecture in high-level vision task. Vision transformer (ViT) proposed by Dosovitskiy et al. [31] apply pure transformer and treats image patches as sequences directly, which achieved state-of-the-art performance on image recognition benchmarks. Now transformer are extended to more vision...
tasks including image processing [32], segmentation [33], [34], pose estimation [35], [36], etc. More recently, there has been some methods proposed to implement the person re-ID task. For example, TransReID [37] uses the ViT, a pure transformer-based model with image patches as the input, to extract a global feature using a lot of transformer encoders which require much more computations and pre-training data than CNNs. Another typical work PAT [38] uses CNNs as the backbone before a transformer encoder to further extract features. It additionally introduces a transformer decoder to mask part features using cross-attention with a set of learnable part prototypes. However, the part prototypes are hard to automatically learn focusing on human parts accurately by end-to-end especially when the occlusions are insufficient in the training data.

Inspired by DETR, We extend transformer to occluded re-ID tasks. Occluded re-ID requires separating, extracting, and matching the features of visible human parts under the interference of occlusions. Since CNN cannot separate features into parts directly, most existing methods utilize an extra pretrained pose estimation or part segmentation model to separate and extract part features while it requires a lot of computation and may suffer poor performance without human parts annotation in the re-ID dataset. However, transformer can naturally separate the feature map into a series of part features with the architecture of encoder-decoder. Specifically, the encoder models the global correlation between feature elements and aggregates them utilizing their similarities, and the decoder transforms the features into N output embeddings by N learnable queries according to the cross attention between the aggregated features and the queries. Therefore, every output embedding represents a semantic component, which can automatically learn to decode features into adaptive parts for better matching.

C. Contrast Feature Learning

Since contrast feature learning methods have been studied for learning discriminative representations, our approach is different from them in goals and practices. Specifically, existing related works can be categorized into two lines. One line of methods, namely contrastive self-supervised learning, such as MoCo [39] and SimCLR [40], aims to learn instance robust features between an image sample and its transformed version sample. However, our approach aims to train occlusion independent features by learning contrast information of relevant and irrelevant features within every training sample. Another line of methods aim to train two contrastive branches with one branch learning class-relevant features and the other learning class-irrelevant features, such as SNR [41] and DAAL [42], which are more related to our approach but also different in two aspects. First is the contrast level and operation. Specifically, SNR and DAAL learn contrastive attention maps for the spatial level or channel level to filter out discriminative information directly by multiplication operation, while our approach learned contrastive transformer decoder queries to decode different features. The second is the feature training manner. For the irrelevant feature, DAAL used a classifier in the ID-irrelevant branch to predict the extra domain styles accurately. SNR trained the ID-irrelevant feature contrarily to closer to ID-negative samples and farther from ID-positive samples without considering any other information. Different from both of them, our approach contained a corresponding sampling strategy that learned with a sample triplet consisting of a positive ID with the negative occlusion and a negative ID with the positive occlusion, where the relevant and irrelevant features were learned by the ID and occlusion reversely. The construction of triplets not only helps better isolate the ID-irrelevant (occlusion) feature but also makes contrast feature learning more interpretable.

III. PROPOSED METHOD

In this section, we firstly introduce the architecture of the proposed DRL-Net in Section III-A, which consists of a CNN and a Transformer. We then elaborate the designed contrast feature learning strategy for the DRL-Net in Section III-C, which suppresses occlusion interference by separation of occlusion features and discriminative ID features. In Section III-D, we explain the training and inference strategies in details. An overview of our method is shown in Fig. 2.

A. Semantic Representation Extraction and Disentanglement

1) Feature Extractor: Our feature extractor contains a CNN backbone and encoder-decoder layers, in order to extract compact representations and generate features of semantic component accordingly.

For a person image x, the CNN backbone generates feature maps v = CNN(x) ∈ R^{C×H×W}, where C, H, W denote the channel dimension, height and width of the feature maps respectively. With the non-linear activation function σ(·), we obtain the activated feature maps a = σ(v) ∈ R^{C×H×W}. A 1 × 1 convolution layer is followed to generate the new feature maps g ∈ R^{d×H×W}, where d is smaller than C to reduce the computation complexity of transformer. In order to construct the sequence form that transformers expect, we flatten the tensor along the last two spatial dimensions and finally get the g ∈ R^{d×HW}.

The encoder-decoder layers in our feature extractor follow the standard architecture of the transformer. We apply learnable positional encodings to encode spatial information and add it to the input of each encoder attention layer. The position encoding p encodes the position information of the i_{th} pixel of the input flattened feature maps g, which is important for the transformer encoder to encode spatial information. There are two choices of positional encodings, learned and fixed [30], [43]. We used learned positional encodings which are the learnable embeddings with the same dimension of the input, so the positional encodings can be summed with the input embedding. We assume each image can be separated into several semantic components, such as the upper body, the legs and the occlusion. To produce features of the semantic components, we define N_s semantic preferences object queries by a set of learnable input embeddings for decoder layers, each of which extracts features of a specific semantic component. The framework consists of N_{q} − 1 ID relevant object queries for the semantic human parts and one ID irrelevant object query for the occlusion part in the image. Utilizing the N_{q} semantic preferences object

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
Fig. 2. The framework of the proposed DRL-Net: DRL-Net consists of three components. The first component is occluded sample augmentation that synthesizes person images by inserting various obstacles. The second component is semantic representation extraction and disentanglement that disentangles the representation of undefined semantic components into ID-relevant features and ID-irrelevant features under the guidance of semantic preference object queries. The third component is contrast feature learning that isolates ID-irrelevant features from global representation by optimizing features in both ID embedding space and occlusion embedding space reversely. Additionally, a decorrelation constraint is imposed over the object queries of the decoder to force them to focus on non-overlapped semantic parts.

queries, denoted as $Q = [q_0, \ldots, q_{N_q-1}, q_m]^\top$, $Q \in \mathbb{R}^{N_q \times d}$, the transformer decoder processes the output of the encoder, separates and transforms them into $N_q$ representation features with each feature corresponding to a semantic component according to a query. These representation features of undefined semantic components output by the decoder are denoted as $F = [f_0, \ldots, f_{N_q-1}, f_m]^\top$, $F \in \mathbb{R}^{N_q \times d}$. Furthermore, features of semantic human parts generated by $N_q - 1$ ID relevant queries are concatenated as the ID relevant feature $f = concat([f_0, \ldots, f_{N_q-1}]) \in \mathbb{R}^{(N_q-1) \times d}$, and the feature of occlusions generated by the ID irrelevant query is considered as the ID irrelevant feature $\tilde{f} = f_m \in \mathbb{R}^d$, which will be utilized to reduce the interference of occlusions and noises.

We adopt cross entropy loss as identity loss to supervise the learning of feature extractor, and label smoothing is used to prevent the model from overfitting training IDs, which is defined as:

$$
\mathcal{L}_{cc} = - \sum_{n=1}^{N} \sum_{m=1}^{M} \sigma_m \log \mathcal{P}_m (f_n),
$$

$$
\sigma_m = \begin{cases} 1 - \epsilon + \frac{\epsilon}{M} & \text{if } m = y_n \\ \frac{\epsilon}{M} & \text{otherwise}, \end{cases}
$$

where $N$ is the number of the training samples, $M$ is the person identity number of the training set, $\mathcal{P}_m (f_n)$ is the predicted probability of feature $f_n$ belonging to identity $m$, and $y_n$ is the ground-truth label of $f_n$. $\sigma_m$ is the smoothing label according to the label $y_n$, and $\epsilon$ is a small constant and set to be 0.1.

2) Object Query Decorrelation Constraint: To extract semantic-aligned features without external supervising, we expect features decoded from different object queries represent different semantic components. We propose object query decorrelation constraint to make object queries orthogonal with each other. Given the set of object queries $Q \in \mathbb{R}^{N_q \times d}$ extracted from person images, the object query decorrelation constraint loss is computed using the following formula:

$$
\mathcal{L}_o = \alpha \sum_{n=0}^{N_q} \sum_{m \neq n} \text{abs} \left( \langle q_n, q_m \rangle \bigg/ \|q_n\| \|q_m\| \right),
$$

where $\text{abs} (\cdot)$ denotes the absolute value function, $\langle \cdot, \cdot \rangle$ denotes the inner product, and $\alpha$ is the penalty factor of decorrelation constraint loss.

The proposed decorrelation constraint is imposed over different object queries to force them to focus on respective semantic components with few overlaps, which helps the transformer better separate and localize the representation of different semantic components.

B. Occluded Sample Augmentation (OSA)

Occluded Sample Augmentation is a data augmentation strategy for our semantic preferences guided contrast feature learning. The limited number of occluded samples in training data often leads to the low diversity of occluded samples in each training batch, which makes the re-ID model sensitive to occlusions. To address this issues, we employ OSA to augment person
images which can preserve the person identities while generating new person images contains multiple obstacles. We first select different obstacles appearing in the train set as obstacle set $X_{\text{obstacle}}$. We selected the obstacles which cover the pedestrian appearance in the training set, maximized the diversity of the species of those obstacles, and ensured they did not show up from the testing set. Crop these obstacles out of the pedestrian images and give them a small range of size and location random change. During the training stage, we randomly selected $k$ obstacles from the $X_{\text{obstacle}}$ to synthesize augmented samples for each training batch. Specifically, given an image batch $B$ and random $k$ obstacles $[o_1, \ldots, o_k] \in X_{\text{obstacle}}$, for each $x_i \in B$ with label $y_i$, we generate augmented image $[x_{i,1}, \ldots, x_{i,k}]$ with label $y_i$ which occluded by $k$ obstacles. In this way, the sample number in each batch increase by a factor of $k$. The augmented images together with original images are used for contrast feature learning (CFL). The benefits of employing OSA can be further demonstrated by introducing CFL.

C. Contrast Feature Learning (CFL)

We proposed semantic preferences guided contrast feature learning to expect semantic components generated by object queries to focus on body parts without the disturbance of occlusions. More specifically, we construct contrast triplets for given person image $x_n$ with the help of OSA, consisting of $x_n$ itself as the anchor, a positive instance with the same ID but different obstacles, and a negative one with different IDs but the same obstacle. The triplet loss with contrast triplets is defined by:

$$
L_{\text{tri}} = \sum_{n=1}^{N} \delta + D(f_n, f_{n+}) - D(f_n, f_{n-})_+, \tag{3}
$$

where $f_n$ denotes the ID relevant features of image $x_n$, and $f_{n+}, f_{n-}$ denote the ID relevant features belonging to the same or different person with $x_n$ respectively. $D(\cdot, \cdot)$ is the distance function between features and $\delta$ is a margin parameter.

Furthermore, we proposed reverse triplet loss to make ID irrelevant features focus on occlusions or noises. We reverse the positive instances and negatives in contrast triplets to guide occlusion object query extract occlusion semantic components in images. The reverse triplet loss is defined by:

$$
L_{r\text{tri}} = \sum_{n=1}^{N} \delta + D(\bar{f}_n, \bar{f}_{n-}) - D(\bar{f}_n, \bar{f}_{n+})_+, \tag{4}
$$

where $\bar{f}_n$ denotes the ID irrelevant feature representations of image $x_n$, and $\bar{f}_{n+}, \bar{f}_{n-}$ denote the ID irrelevant features belonging to the same or different person with $x_n$ respectively. For the reverse triplet loss $L_{r\text{tri}}$, in order to learn discrimination of occlusions, the irrelevant feature of the anchor should be more close to the sample with the positive occlusion in the triplet even though they have different person identities (negative identity). Under this constraint, the irrelevant feature will activate on the occlusion areas in the image. With the proposed decorrelation constraint and CFL, we force occluded semantic components only extracted by ID irrelevant object query, making human semantic components extracted by ID relevant queries free from occlusions.

D. Training and Inference

The training and inference process of the proposed DRL-Net is shown in Algorithm 1. Before the training stage, the obstacle set is constructed by obtaining obstacles from training images. In the occluded sample augmentation stage, given a mini-batch of images for training, we generate augmented samples with random obstacles to obtaining positive pairs and negatives. The entire feature extractor containing convolutional layers and encoder-decoder layers are trained together with the overall loss. The overall loss is therefore calculated as:

$$
L = L_{\text{ce}} + L_o + L_{\text{tri}} + \lambda L_{r\text{tri}}, \tag{5}
$$

where $\lambda$ is the scale factor of reverse triplet loss, and the scale factors of others are set to be 1.

In the inference stage, query and gallery images are the input to feature extractor without augmentation, and we utilize ID relevant feature $f$ to compute the distance between query and gallery images, ignoring the ID irrelevant feature $\bar{f}$. 

---

Algorithm 1: Proposed DRL-Net.

**Input:** Training/query/gallery: $X_{\text{train}}, X_{\text{query}}, X_{\text{gallery}}$

**Output:** Distance Matrix $D$

1: %Data preparation
2: Obtain the obstacle set $X_{\text{obstacle}}$ from $X_{\text{train}}$
3: %Training stage
4: Initialize the CNN network parameters $\Theta$
5: for each mini-batch $B \subset X_{\text{train}}$ do
6: Create set $B' = \emptyset$
7: Random select obstacle $[o_1, \ldots, o_k] \in X_{\text{obstacle}}$
8: for each $x_i \in B$ do
9: Generate augmented sample $[x_{i,1}, \ldots, x_{i,k}]$ by $x_i$
10: Adding $[x_i, x_{i,1}, \ldots, x_{i,k}]$ to set $B'$
11: end for
12: Extract ID relevant feature $f_i$ and irrelevant feature $\bar{f}_i$ of each $x_i \in B'$ using DRL-Net.
13: Calculate $L_{\text{ce}}, L_o, L_{\text{tri}}, L_{r\text{tri}}$ by (1), (2), (3), and (4).
14: Optimize CNN parameters $\Theta$ according to (5).
15: end for
16: %Inference stage
17: for each $x_q \in X_{\text{query}}, x_g \in X_{\text{gallery}}$ do
18: Extract $f_q, f_g$ of $x_q, x_g$ respectively using DRL-Net.
19: Calculate $D(f_q, f_g)$ by cosine distance metric.
20: end for
21: return $D$
IV. EXPERIMENTS

A. Datasets and Evaluation Metrics

The experiments are conducted on three person ReID datasets, including one occluded re-ID dataset Occluded-DukeMTMC and two widely used Holistic re-ID datasets MSMT1 and Market-1501.

Occluded-DukeMTMC [12] is a split of DukeMTMC-reID [44] which keeps occluded images and removes some overlap images. It contains 15,618 training images, 17,661 gallery images, and 2,210 occluded query images, which is by far the largest occluded re-ID datasets. The experiments on this dataset follow the standard setting [12] and the training, query, and gallery sets contain 9%, 100%, and 10% occluded images, respectively. Market-1501 [45] consists of 32,668 images of 1,501 identities captured by 6 camera views. Following the standard setting [45], the whole dataset is divided into a training set containing 12,936 images of 751 identities and a testing set containing 2,228 query images and 17,661 gallery images.

Evaluation metric: We adopt Cumulative Matching Characteristic (CMC) curve and mean average precision (mAP) for evaluations. All experiments are conducted in the single query mode and don’t use Re-Ranking to further refine the results.

B. Implementation Details

Consistent implementation details are adopted for all datasets in our experiments.

Data preprocessing: All person images are resized to 256 × 128 in both training and inference stages. The training images are augmented with random horizontal flipping, random cropping and random erasing [47] with a probability of 0.5. We construct an occlusion set by fetch obstacles which hardly appeared in test images from train set. All train images are copied and synthesized with random obstacles from our occlusion set in occluded sample augmentation stage.

Backbones: We adopt ResNet-50 [48] as the backbone network. Following the setting of most ReID methods [49], the last spatial down-sampling operation in ResNet-50 is removed to increase the spatial size of the feature map. In this case, the size of the feature map is 2048 × 16 × 8. The hidden dimension d is set to 256. The transformer layers are same with DETR and initialized with Xavier init [50]. The numbers of encoder layers, decoder layers and multi-head attention are set to 2, 2, 8 respectively. The cosine distance is used to measure the distance between query and gallery images.

C. Comparison with the State-of-the-Art

We compare our method with state-of-the-art methods for both occluded and holistic person re-ID tasks in Table I and Table II, respectively. The backbones of compared methods are ResNet-50 or modified ResNet-50 by using branches, attentions or different convolution operations.

Results on Occluded-DukeMTMC: The comparison results over dataset Occluded-DukeMTMC are shown in Table I. There are four mainstream types of occluded re-ID methods are compared: holistic re-ID methods without designing modules for occlusions (DIM [52], Part Aligned [53], HACNN [54], Adver Occluded [55] and PCB [23]), occluded re-ID methods with external cues (Part Bilinear [56], FD-GAN [57], PGFA [12] and HONet [17]), occluded re-ID methods based on part-to-part matching (DSR [19], SFR [58] and MoS [59]), and re-ID methods based on transformer. Our DRL-Net achieves 65.8% Rank-1 accuracy, 80.4% Rank-5 accuracy, 85.2% Rank-10 accuracy, and 53.9% mAP, which outperforms all types of methods using the same backbone of ResNet-50 by a large margin. It also achieves comparable performance with TransReID which uses a stronger backbone of ViT. The superior performance of DRL-Net is summarized into three aspects. First, the transformer layers enhance the representation capacity of CNN backbones. Second, the introduction

TABLE I

| Methods                | Rank-1 | Rank-5 | Rank-10 | mAP  |
|------------------------|--------|--------|---------|------|
| DIM (ArXiv 17)         | 21.5   | 36.1   | 42.8    | 14.4 |
| Part Aligned (ICCV 17) | 28.8   | 44.6   | 51.0    | 20.2 |
| HACNN (CVPR 18)        | 34.4   | 51.9   | 59.4    | 26.0 |
| Adver Occluded (CVPR 18)| 44.5  | -      | -       | -    |
| PCB (ECCV 18)          | 42.6   | 57.1   | 62.9    | 33.7 |
| Part Bilinear (ECCV 18)| 36.9   | -      | -       | -    |
| FD-GAN (NIPS 18)       | 40.8   | -      | -       | -    |
| PGFA (ICCV 19)         | 51.4   | 68.6   | 74.9    | 37.3 |
| HONet (CVPR 20)        | 55.1   | -      | -       | 43.8 |
| DSR (CVPR 18)          | 40.8   | 58.2   | 65.2    | 30.4 |
| SFR (ArXiv 18)         | 42.3   | 60.3   | 67.3    | 32.0 |
| ISP* (ECCV 20)         | 62.8   | 78.1   | 82.9    | 52.3 |
| MoS (AAAI 21)          | 61.0   | 74.4   | 79.1    | 49.2 |
| PAT (CVPR 21)          | 64.5   | -      | -       | 53.6 |
| TransReID* (ECCV 21)   | 66.4   | -      | -       | 59.2 |
| DRL-Net (Ours)         | 65.8   | 80.4   | 85.2    | 53.9 |

Optimization: The CNN backbone network is pretrained over ImageNet [51]. Adam optimizer is adopted and we warm up the model for 10 epochs with a linearly growing learning rate from 3.5 × 10⁻⁵ to 3.5 × 10⁻⁴. The learning rate is decreased by a factor of 0.1 at 40th and 70th epoch. The batch size is set to 32 with 4 images per ID. The margins for Ltri and Lstr are set to 0.1 and 0.3, respectively.
of object queries and decorrelation constraint gives our method the ability to learn disentangled representation implicitly without external cues. Third, the novel metric learning CFL with synthesized images efficiently weakens the interference of occlusions and noises.

\textbf{Results on Market-1501 and DukeMTMC:} The comparison results over holistic re-ID datasets including Market-1501 and DukeMTMC are shown in Table II. Three types of holistic re-ID methods are considered for comparison: methods based on global features (IANet [60], MVPM [61], DMMIL [62], SFT [63], VCFL [64], MoS [59]), methods using part features (PCB, PCB+RPP [23], AlignedReID [24], DSR [19] and VPM [20]), and methods using external cues including human-parsing based (SPReID [15] and MGCAM [13]), attribute information based (AANet [66]) and human pose based (Pose-transfer [67], PSE [68], PGFA [12] and HONet [17]). Our method is specially designed for a more challenging scenario of occluded ReID. Specifically, the two datasets contain relatively perfect images to identify without the interference of occlusions, and thus make our approach hard to give play to advantages. Though DRL-Net is not proposed for the holistic re-ID task, it still guarantees a comparable performance with most of holistic re-ID methods, indicating its robustness and generalization.

\textbf{Results on MSMT17.} Since MSMT17 is released recently, hence there are only a few methods that report on this dataset, including MVMP [61], SFT [63], DG-Net [69], IANet [60], Circle [65] and Circle + MGN [65]. Table III shows the comparison results. DRL-Net achieves outstanding performance in all evaluation metrics.

\textbf{Cross dataset experiments:} We also experiment with DRL-Net in a cross dataset setting. Since the method requires occluded images for training, the model is trained in Occluded-DukeMTMC and tested on DukeMTMC, Market1501 and MSMT17. As Table VIII shows, DRL-Net performs relatively

\begin{table}[h]
\centering
\caption{Comparison over Datasets Market-1501 and DukeMTMC shows DRL-Net can be generalized to holistic re-ID with superior performance. The compared methods are grouped into three categories: global feature based, part feature based and external cues based.}
\begin{tabular}{|l|c|c|c|}
\hline
\textbf{Methods} & \textbf{Market-1501} & \textbf{DukeMTMC} & \\
\hline
IANet (CVPR 19) & 94.4 & 83.1 & 87.1 & 73.4 \\
MVPM (ICCV 19) & 91.4 & 80.5 & 83.4 & 70.0 \\
DMMIL (ICCV 19) & 93.5 & 81.6 & 85.9 & 73.7 \\
SFT (ICCV 19) & 93.4 & 82.7 & 86.9 & 73.2 \\
VCFL (ICCV 19) & 89.3 & 74.5 & - & - \\
Circle (CVPR 20) & 94.2 & 84.9 & - & - \\
\hline
PCB(ECCV 18) & 92.3 & 77.4 & 81.8 & 66.1 \\
PCB+RPP (ECCV 18) & 93.8 & 81.6 & 83.3 & 69.2 \\
AlignedReID(AdaNet 18) & 91.8 & 79.3 & - & - \\
DSR (CVPR 18) & 83.6 & 64.3 & - & - \\
VPM (CVPR 19) & 93.0 & 80.8 & 83.6 & 72.6 \\
\hline
\textbf{DRL-Net (Ours)} & \textbf{94.7} & \textbf{86.9} & \textbf{88.1} & \textbf{76.6} \\
\hline
\end{tabular}
\end{table}
well on DukeMTMC-reID due to the smaller domain gap and not well on the other datasets because of the larger domain gaps. Note that DRL-Net aims to solve the occlusion problem in person re-ID, while the performance drop is mainly attributed to the domain discrepancy between datasets.

**D. Ablation Study**

In this section, we conducted extensive ablation studies to investigate the effectiveness of each component of DRL-Net. We used ResNet-50 as backbone and performed ablation experiments over Occluded-DukeMTMC and the holistic re-ID datasets. Table V and Table IV show experimental results over Occluded-DukeMTMC and holistic re-ID datasets.

**Effectiveness of the proposed Transformer Architecture:** We first study the effect of proposed transformer-based feature extractor which is denoted as baseline+T by removing the CFL and OSA in the framework. The Baseline model which directly uses ResNet-50 as feature extractor and the baseline+T model are both trained by original triplet loss as well as identity loss. As shown in the first two rows of Table V, consistent improvements are achieved on all four evaluation metrics. This indicates that the transformer has a strong ability for feature extracting and disentangling through conducting the global reasoning to further combine the features, which helps to handle the occlusion challenge effectively. The benefits of employing transformer architecture can be further demonstrated by introducing other relevant designs and operations.

**Effectiveness of the proposed OSA:** We evaluate the occluded sample augmentation as described in Section III-B. For this experiment, we design a network baseline+T+OSA that just incorporates the occluded sample augmentation into the baseline+T and maintain the training strategy. As shown in Table V, occluded sample augmentation can improve the re-ID performance on CMC Rank-1/5/10. The improvement can be explained by the effectiveness of the augmented samples that increases the diversity of occluded training samples. On the other hand, due to the gap between the synthesized occluded images and the real images, it suffers a slight decrease in mAP when simply incorporating the occluded sample augmentation.

**Effectiveness of the proposed CFL:** We further evaluate the contrast feature learning component as described in Section III-B. For this experiment, We incorporate contrast feature learning into the baseline+T+OSA as described in the previous subsection and we denote it as baseline+T+OSA+CFL. As Table V shows, the incorporation of contrast feature learning significantly improves the re-ID performance beyond baseline+T+OSA. The baseline+T+OSA+CFL achieves a rank-1 accuracy of 65.8% and an mAP of 53.9% which outperforms the corresponding baseline+T+OSA by 5.3% and 5.7%, respectively. The effectiveness of the CFL can be largely attributed to the separation of occlusions feature and discriminative ID-relevant features, which is crucial to eliminate interference from occlusions for occluded person re-ID.

The ablation studies show that the proposed DRL-Net outperforms the Baseline by 14.8% in Rank-1 accuracy and 10.1% in mAP while working with the occluded sample augmentation and contrast feature learning. This demonstrates that the three components complement each other in achieving better occluded re-ID performance.

Further to investigate the robustness and generalization of DRL-Net, we conduct the additional ablation studies over holistic re-ID datasets including Market-1501, Duke-MTMC and MSMT17. As shown in Table IV, we compare our method with several models: (1st line) B. uses ResNet-50 as the backbone and the cross entropy loss as well as original triplet loss are imposed on it. (2nd line) B.+T+w/o ,L o uses transformer-based feature extractor as described in Section III-A. We removing the proposed object query decorrelation constraint L o, the occluded sample augmentation and contrast feature learning in the framework. (3rd line) B.+T+w/ ,Lo, o adds additional object query decorrelation constraint L o compared to B.+T+w/ ,Lo, o. (4th line) B.+T+w/ ,Lo, o+OSA directly adopts the occluded sample augmentation without contrast feature learning. The performance degradation that simply utilize OSA can be attributed to the few occluded sample in the testing set for holistic re-ID task and the gap between the synthesized occluded images and the real images. (5th line) DRL-Net (overall) is the final version which incorporates contrast feature learning into the B.+T+w/ ,Lo, o+OSA. The Table IV shows the separate contribution from each components of DRL-Net and their generalization to non-occluded scenario.

**E. Parameter Analysis**

**The Number of Transformer Layers:** The impressive performance that transformer achieved can largely contribute to its self-attention mechanism, with which transformer can globally model relations between feature representations of different semantic components. To evaluate the importance of self-attention mechanism, we conduct experiments by changing the number of encoder-decoder layers N l as shown in Table VI. We observe that when the N l is set to 2, the best re-ID performance is achieved, and then the improvement brought by transformer diminishes as depth increases. We think it is because the re-ID task utilizes the lower-resolution representations throughout the network than other high-level vision tasks (e.g. object detection). Moreover, the scale of the re-ID datasets is relatively small and the image contents in datasets are simple, which makes the cross-correlations between the output elements of the decoder are easy to compute.

**The Number of Semantic Preferences Object Queries:** Intuitively, the number of semantic preferences object queries N q determines the granularity of the semantic components. We perform the quantitative ablation studies to find the most suitable N q. As detailed in Table VII, the performance of DRL-Net is robust to different N q. We can observe that as the N q increases, the re-ID performance is continuously improved, but the inference cost also increases correspondingly. To balance performance and cost, we finally set N q to 9.

**The Robustness of Parameters λ and ω:** We studied hyper parameters in DRL-Net by setting it to different values and checking the person Re-ID performance. Fig. 4 shows the experimental results on Occluded-DukeMTMC dataset. We first analyze the influence of λ in Fig. 4(a), the scale factor λ in (5) is theAuthorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
balancing weight of contrast feature learning (CFL). With $\lambda$ increasing, the Rank-1/mAP is improved by 4.5%/2.6% ($\lambda = 1.0$), which means the CFL module now is beneficial for learning better occlusion robust re-ID features. Continuing to increase $\lambda$, the performance is degraded because the weights for human parts feature embedding and the position embedding are weakened.

Then we analyze the effect of penalty factor $\alpha$ on baseline+T model as shown in Fig. 4(b). The penalty factor $\alpha$ in (2) will affect correlation among object queries in decoder. Experiments show that baseline+T performs best when $\alpha = 1.0$. Using a smaller $\alpha$ will suppress the value of $L_\alpha$, lower the decorrelation ability for object queries. On the other hand, $\alpha$ should not be very large for preserving the feature representation capability of the model. Experimental results both in (a) and (b) show that DRL-Net performs stably and is tolerant to the change of the parameters.

**F. Effectiveness of Each Loss**

The cross entropy loss $L_{ce}$: trains re-ID models as a classification problem by predicting the identity for every image, which is a common re-ID baseline introduced by IDE [1]. Removing the $L_{ce}$, the performance significantly drops by 19.2% Rank-1 score and 18.4% mAP, indicating that the $L_{ce}$ can provide a basic ability for discriminative feature learning.

The object query decorrelation constraint loss $L_{o}$: aims to learn diverse discrimination for the transformer decoder by making object queries focus on non-overlapped image areas. To this end, it makes object queries orthogonal with each other by minimizing their cosine similarity. Compared with the overall result, the performance drops 2.1% Rank-1 score and 5.7% mAP without $L_o$ due to the redundant information and the lack of complementary capabilities of object queries.

The triplet loss $L_{tri}$: trains re-ID models as a rank problem by ensuring that an image of a specific person is closer to all other images of the same person than to any images of other persons. Without $L_{tri}$, the performance drops 0.9% mAP, which indicates its effectiveness for providing more discriminative capability of re-ID.

The reverse triplet loss $L_{rtri}$: aims to filter out the disturbing occlusion information from the image features for better identification. To this end, it makes the last object query focus on the occlusion areas by training its corresponding features to identify the categories of the obstacle, and thus other object queries may focus on the human body because they are orthogonal with the last query. Specifically, $L_{rtri}$ learns to identify obstacles in a reverse triplet manner, where the positive person is with a negative occlusion and the negative person is with a positive occlusion, making features of the same occlusion close to each other and features of different occlusions far from each other. As shown in Table IX, without $L_{rtri}$ the performance drops by 5.3% Rank-1 and 5.7% mAP, validating its effectiveness for alleviating the impacts of occlusion.

**G. Visualization**

**Qualitative Results.** We demonstrate how DRL-Net overcomes the occlusion constraint by providing several samples of
This paper proposes a novel alignment-free method DRL-Net that handles occluded re-ID through disentangled representation learning. Leveraging transformer architectures, DRL-Net performs global reasoning based on the interrelation of undefined semantic components, which allows feature disentanglement without any supervision of part correspondences. Furthermore, to better eliminate the interference of occlusion noises, we design a contrast feature learning technique to encourage the separation of occlusions feature and ID-relevant features. Extensive experimental evaluations on several benchmarks demonstrate that DRL-Net achieves superior re-ID performance consistently.

TABLE IX
EVALUATION THE IMPORTANCE OF EACH COMPONENT OF THE OVERALL LOSS FUNCTION OVER THE OCCLUDED-DUREMTMC

| $L_{ce}$ | $L_{tv1}$ | $L_{o}$ | $L_{tv2}$ | Rank-1 mAP |
|---------|----------|--------|----------|-----------|
| ✓       | ✓        | ✓      | ✓        | 65.8      | 53.9      |
| ✓       | ✓        | ✓      | ✓        | 46.6      | 35.5      |
| ✓       | ✓        | ✓      | ✓        | 65.3      | 53.0      |
| ✓       | ✓        | ✓      | ✓        | 63.7      | 48.2      |
| ✓       | ✓        | ✓      | ✓        | 60.5      | 48.2      |

Fig. 5. Visualization of the decoder attention of ID-irrelevant object queries: For each of the eight image pairs, the left shows the original person image and the right shows the heat map of ID irrelevant object query.

V. Conclusion

This paper proposes a novel alignment-free method DRL-Net that handles occluded re-ID through disentangled representation learning. Leveraging transformer architectures, DRL-Net performs global reasoning based on the interrelation of undefined semantic components, which allows feature disentanglement without any supervision of part correspondences. Furthermore, to better eliminate the interference of occlusion noises, we design a contrast feature learning technique to encourage the separation of occlusions feature and ID-relevant features. Extensive experimental evaluations on several benchmarks demonstrate that DRL-Net achieves superior re-ID performance consistently.

REFERENCES

[1] L. Zheng, Y. Yang, and A. G. Hauptmann, “Person re-identification: Past, present and future,” 2016, arXiv:1610.02984.
[2] S. Gong, M. Cristani, C. C. Loy, and T. M. Hospedales, “The re-identification challenge,” in Person Re-Identification. Berlin, Germany: Springer, 2014.
[3] S. Zhou, J. Wang, R. Shi, Q. Hou, Y. Gong, and N. Zheng, “Large margin learning in set-to-set similarity comparison for person reidentification,” IEEE Trans. Multimedia, vol. 20, no. 3, pp. 593–604, Mar. 2018.
[4] Z. Wang, J. Jiang, Y. Yu, and S. Satoh, “Incremental re-identification by cross-direction and cross-ranking adaption,” IEEE Trans. Multimedia, vol. 21, no. 9, pp. 2376–2386, Sep. 2019.
[5] H. Luo et al., “A strong baseline and batch normalization neck for deep person re-identification,” IEEE Trans. Multimedia, vol. 22, no. 10, pp. 2597–2609, Oct. 2019.
[6] M. Cao, C. Chen, H. Dou, X. Hu, S. Peng, and A. Kuijper, “Progressive bilateral-context driven model for post-processing person reidentification,” IEEE Trans. Multimedia, vol. 23, pp. 1239–1251, 2020.
[7] Z. Wang et al., “Zero-shot person re-identification via cross-view consistency,” IEEE Trans. Multimedia, vol. 18, no. 2, pp. 260–272, Feb. 2016.
[8] Z. Zeng, Z. Wang, Z. Yang, Z. Zheng, Y.-Y. Chuang, and S. Satoh, “Illumination-adaptive person re-identification,” IEEE Trans. Multimedia, vol. 22, no. 12, pp. 3064–3074, Dec. 2020.
[9] Y. Wang, R. Hu, C. Liang, C. Zhang, and Q. Leng, “Camera compensation using a feature projection matrix for person reidentification,” IEEE Trans. Circuits Syst. Video Technol., vol. 24, no. 8, pp. 1350–1361, Aug. 2014.
[10] L. Wei, S. Zhang, H. Yao, W. Gao, and Q. Tian, “Glad: Global-local-alignment descriptor for scalable person re-identification,” IEEE Trans. Multimedia, vol. 21, no. 4, pp. 986–999, Apr. 2019.
[11] J. Zhuo, Z. Chen, J. Lai, and G. Wang, “Occluded person re-identification,” in Proc. IEEE Conf. Multimedia Expo, 2018, pp. 1–5.
[12] J. Miao, Y. Wu, P. Liu, Y. Ding, and Y. Yang, “Pose-guided feature alignment for occluded person re-identification,” in Proc. IEEE Int. Conf. Comput. Vis., 2019, pp. 542–551.
[13] C. Song, Y. Huang, W. Ouyang, and L. Wang, “Mask-guided contrastive attention model for person re-identification,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 1179–1188.
[14] L. He, Y. Wang, W. Liu, H. Zhao, Z. Sun, and J. Feng, “Foreground-aware pyramid reconstruction for alignment-free occluded person re-identification,” in Proc. IEEE Int. Conf. Comput. Vis., 2019, pp. 8450–8459.
[15] M. M. Kalayeh, E. Basaran, M. Gökmen, M. E. Kamasak, and M. Shah, “Human semantic parsing for person re-identification,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 1062–1071.
[16] S. Gao, J. Wang, H. Lu, and Z. Liu, “Pose-guided visible part matching for occluded person reid,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2020, pp. 1174–11752.
[17] G. Wang et al., “High-order information matters: Learning relation and topology for occluded person re-identification,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2020, pp. 6449–6458.
[18] W.-S. Zheng, X. Li, T. Xiang, S. Liao, J. Lai, and S. Gong, “Partial person re-identification,” in Proc. IEEE Int. Conf. Comput. Vis., 2015, pp. 4678–4686.
[19] L. He, J. Liang, H. Li, and Z. Sun, “Deep spatial feature reconstruction for partial person re-identification: Alignment-free approach,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 7073–7082.

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
Mengxi Jia received the B.S. degree from Central South University, Changsha, China. She is currently working toward the M.S. degree with the School of Electronic and Computer Engineering, Shenzhen Graduate School, Peking University, Shenzhen, China. Her research interests include person re-identification and deep metric learning.

Xinhua Cheng received the B.E. degree in computer science and technology from Sichuan University, Chengdu, China. He is currently working toward the M.S. degree with the School of Electronic and Computer Engineering, Peking University, Shenzhen, China. His research interests include person reidentification, deep learning and computer vision.

Shijian Lu received the Ph.D. degree in electrical and computer engineering from the National University of Singapore, Singapore. He is currently an Assistant Professor with the School of Computer Science and Engineering, Nanyang Technological University, Singapore. He has authored or coauthored more than 100 international refereed journal and conference papers and coauthored more than ten patents in these research areas. His major research interests include image and video analytics, visual intelligence, and machine learning. He is also an Associate Editor for the journal Pattern Recognition. He also was the Program Committee of a number of conferences, the Area Chair of the International Conference on Document Analysis and Recognition (ICDAR) 2017 and 2019, the Senior Program Committee of the International Joint Conferences on Artificial Intelligence (IJCAI) 2018 and 2019.

Jian Zhang (Member, IEEE) received the B.S. degree from the Department of Mathematics, Harbin Institute of Technology (HIT), Harbin, China, in 2007, and the M.Eng. and Ph.D. degrees from the School of Computer Science and Technology, HIT, in 2009 and 2014, respectively. From 2014 to 2018, he was a Post-doctoral Researcher with Peking University, Beijing, China, Hong Kong University of Science and Technology, Hong Kong, and the King Abdullah University of Science and Technology, Thuwal, Saudi Arabia. He is currently an Assistant Professor with the School of Electronic and Computer Engineering, Peking University Shenzhen Graduate School, Shenzhen, China. He has authored or coauthored 80 technical articles in refereed international journals and proceedings. His research interests include intelligent multimedia processing, deep learning, and optimization. He was the recipient of the Best Paper Award at the 2011, IEEE Visual Communications and Image Processing (VCIP) and was the corecipient of the Best Paper Award of 2018 IEEE MultiMedia.