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

Mengxi Jia, Xinhua Cheng, Shijian Lu and Jian Zhang

***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 (Occluded-DukeMTMC, Market1501 and DukeMTMC) show that the DRL-Net achieves superior re-ID performance consistently and outperforms the state-of-the-art by large margins for Occluded-DukeMTMC. Code is available at https://github.com/Anonymous-release-code/DRL-Net.**

***Index Terms*—Person re-identification, representation learning, visual Transformer, Occlusion Scene.**

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], [3], [4], [5], [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.

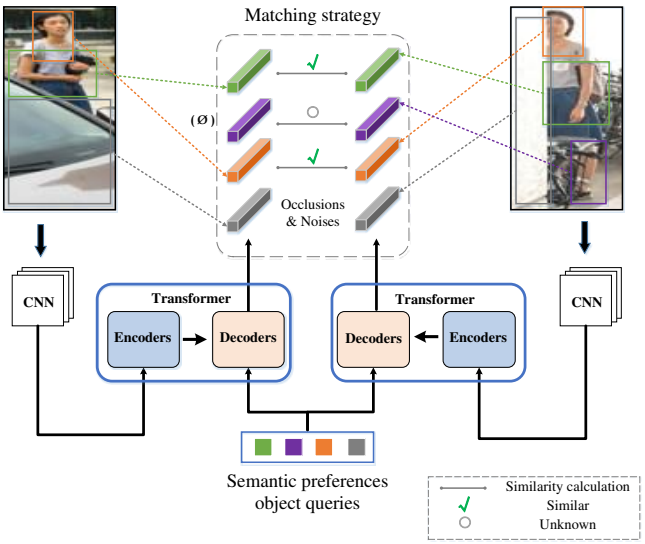

Fig. 1: Illustration of the proposed DRL-Net in occluded person re-ID: Person images often suffer from various occlusions with different visible and invisible body parts which greatly complicate image alignment and similarity computation. The proposed DRL-Net is alignment-free which exploits semantic preferences object queries that guide transformer to disentangle the representation of body parts and eliminate occlusion noises in similarity measurement.

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

Manuscript received July 6, 2021. This work was supported in part by Key-Area Research and Development Program of Guangdong Province (2019B121204008) and National Natural Science Foundation of China (61902009). *(corresponding author: Jian Zhang.)*

Mengxi Jia, Xinhua Cheng and Jian Zhang are with the School of Electronic and Computer Engineering, Shenzhen Graduate School, Peking University, Shenzhen 518055, China. (E-mail: {mxjia, zhangjian.sz}@pku.edu.cn)

Shijian Lu is with the School of Computer Science and Engineering, Nanyang Technological University, Singapore 639798, Singapore. (E-mail: shijian.lu@ntu.edu.sg)

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], [19], [20], [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 the transformer architecture that disentangles CNN features together with positional encoding into ID-relevant features (for person image matching) and ID-irrelevant features (for eliminating occlusion interference). DRL-Net performs global reasoning based on the interrelation of undefined semantic components, which allows feature disentanglement without any supervision of part correspondences and accordingly avoids the complicated and error-prone 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 re-ID 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], [27], [28], [29], researchers applied transformer to various computer vision areas. Carion et al. [30] present 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

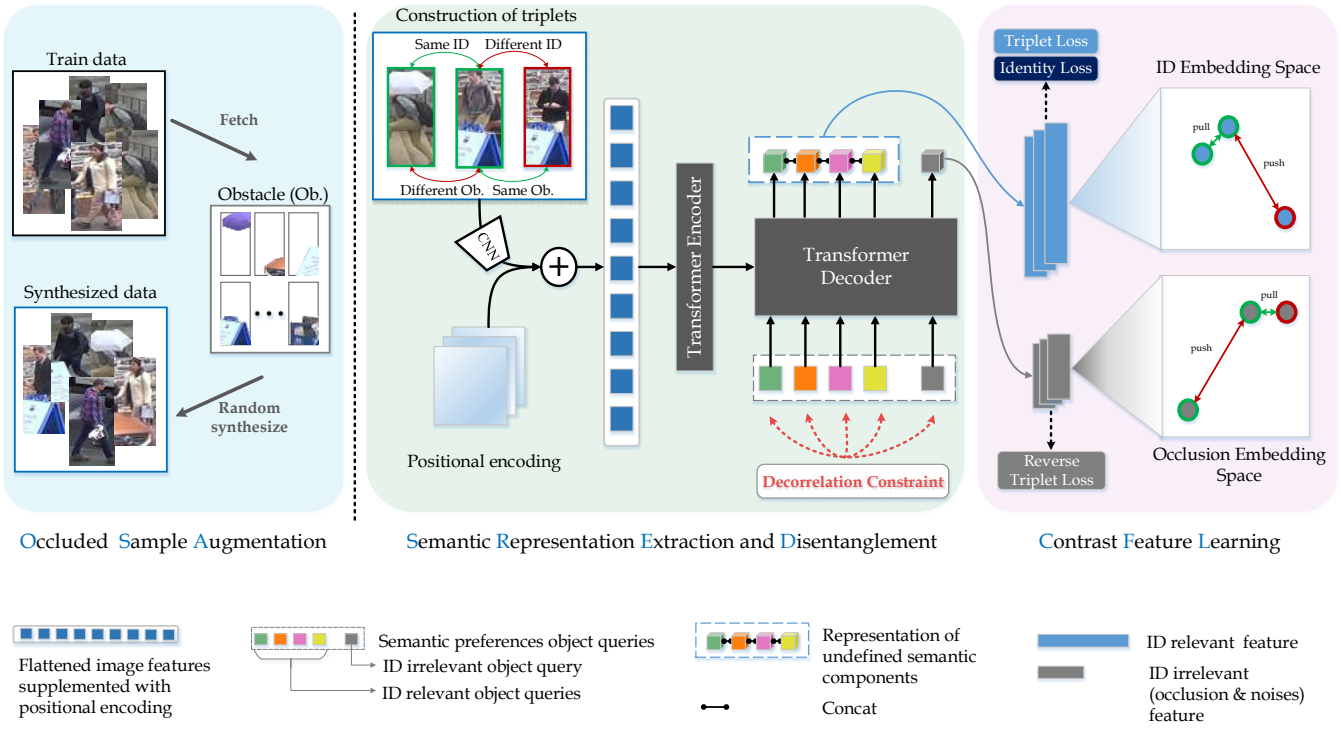

Fig. 2: The framework of the proposed DRL-Net: DRL-Net consists of three components. The first component is **occluded sample augmentation** that synthesizes person images by inserting various obstacles. The second component is **semantic representation extraction and disentanglement** that disentangles 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.

are extended to more vision tasks including image processing [32], segmentation [33], [34], pose estimation [35], [36], etc.

We extend DETR to occluded re-id tasks, using object queries to extract features of semantic components instead of extra prediction or artificial pre-definition. Benefit from the outstanding representation capabilities of transformer architectures, our method achieves promising performance.

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-B, which suppresses occlusion interference by separation of occlusion features and discriminative ID features. In Section III-C, 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 $\mathbf{f} = CNN(x) \in \mathbb{R}^{C\times H\times W}$, where C, H, W denote the channel dimension, height and width of the feature maps respectively. With the non-linear activation function $\sigma(\cdot)$, we obtain the activated feature maps $\mathbf{a} = \sigma(\mathbf{f}) \in \mathbb{R}^{C\times H\times W}$. A 1×1 convolution layer is followed to generate the new feature maps $\mathbf{g} \in \mathbb{R}^{d\times H\times W}$, 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 $\mathbf{g} \in \mathbb{R}^{d\times HW}$.

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. To produce features of semantic components, we define semantic preferences object queries, which are a set of learnable input embeddings for decoder layers. More specifically, there are $N_q - 1$ human semantic object queries and 1 occlusion object query, where N_q is the semantic preferences object query number. The semantic preferences object queries denoted as $\mathbf{Q} = [\mathbf{q}_0, ..., \mathbf{q}_{N_q-1}; \mathbf{q}_o], \mathbf{Q} \in \mathbb{R}^{N_q\times d}$ are different to generate representation features and they are added to the input of

each attention layer. The representation features of undefined semantic components generated by $\mathbf{Q}$ and encoder-decoder layers are denoted as $\mathbf{F} = [\mathbf{f}_0, ..., \mathbf{f}_{N_q-1}; \mathbf{f}_o], \mathbf{F} \in \mathbb{R}^{N_q \times d}$. Features of human semantic components generated by $N_q - 1$ ID relevant queries are concatenated as the ID relevant feature $\mathbf{f} = concat([\mathbf{f}_0, ..., \mathbf{f}_{N_q-1}]) \in \mathbb{R}^{(N_q-1)\cdot d}$, and ID irrelevant feature $\bar{\mathbf{f}} = \mathbf{f}_o \in \mathbb{R}^d$ generated by occlusion query are used 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:

$$\begin{aligned} \mathcal{L}_{ce} &= -\sum_{n=1}^{N}\sum_{m=1}^{M} q_m \log \mathcal{P}_m\left(\mathbf{f}_n\right), \\ q_m &= \begin{cases} 1-\epsilon+\frac{\epsilon}{M} & \text{if } m = y_n \\ \frac{\epsilon}{M} & \text{otherwise,} \end{cases} \end{aligned} \tag{1}$$

where N is the number of training samples, M is the person identity number of the training set, $\mathcal{P}_m(\mathbf{f}_n)$ is the predicted probability of feature $\mathbf{f}_n$ belonging to identity m, and y_n is the ground-truth label of $\mathbf{f}_n$. q_m is the smoothing label according to the label y_n and ϵ 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. Giving the set of object queries $\mathbf{Q}^i \in \mathbb{R}^{N_q \times d}$ extracted from person image i, the object query decorrelation constraint loss is computed using the following formula:

$$\mathcal{L}_o = \alpha \sum_{i=1}^{N}\sum_{n=1}^{N_q}\sum_{m=1}^{N_q} abs(\frac{\left\langle \mathbf{q}_n^i, \mathbf{q}_m^i \right\rangle}{||\mathbf{q}_n^i||\ ||\mathbf{q}_m^i||}), \tag{2}$$

where $abs(\cdot)$ denotes the absolute value function, $\langle \cdot, \cdot \rangle$ denotes the inner product, and α 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. Semantic Preferences guided Contrast Feature Learning

1) 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 $\mathcal{X}_{obstacle}$. During the training stage, we randomly selected k obstacles from the $\mathcal{X}_{obstacle}$ to synthesize augmented samples for each training batch. Specifically, given an image batch $\mathcal{B}$ and random k obstacles $[\mathbf{o}_1, ..., \mathbf{o}_k] \in \mathcal{X}_{obstacle}$, for each $\mathbf{x}_i \in \mathcal{B}$ with label y_i, we generate augmented image $[\mathbf{x}_{i,1}, ..., \mathbf{x}_{i,k}]$ 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.

2) 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 $\mathbf{x}$ with the help of OSA, consisting of $\mathbf{x}$ 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:

$$\mathcal{L}_{tri} = \sum_{n=1}^{N}[\delta + \mathcal{D}(\mathbf{f}_n, \mathbf{f}_{n+}) - \mathcal{D}(\mathbf{f}_n, \mathbf{f}_{n-})]_+, \tag{3}$$

where $\mathbf{f}_n$ denotes the ID relevant features of image $\mathbf{x}_n$, and $\mathbf{f}_{n+}, \mathbf{f}_{n-}$ denote the features belonging to the same or different person with $\mathbf{f}_n$ respectively. $\mathcal{D}(\cdot, \cdot)$ is the distance function between features and δ 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:

$$\mathcal{L}_{rtri} = \sum_{n=1}^{N}[\delta + \mathcal{D}(\bar{\mathbf{f}}_n, \bar{\mathbf{f}}_{n-}) - \mathcal{D}(\bar{\mathbf{f}}_n, \bar{\mathbf{f}}_{n+})]_+, \tag{4}$$

where $\bar{\mathbf{f}}_n$ denotes the ID irrelevant feature representations of image $\mathbf{x}_n$, and $\bar{\mathbf{f}}_{n+}, \bar{\mathbf{f}}_{n-}$ denote the features belonging to the same or different person with $\bar{\mathbf{f}}_n$ respectively. 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.

C. 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:

$$\mathcal{L} = \mathcal{L}_{ce} + \mathcal{L}_o + \mathcal{L}_{tri} + \lambda \mathcal{L}_{rtri}, \tag{5}$$

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

Algorithm 1 Proposed DRL-Net

Input: Training/query/gallery: $\mathcal{X}_{train}$, $\mathcal{X}_{query}$, $\mathcal{X}_{gallery}$
Output: Distance Matrix $\mathcal{D}$

%*Data preparation*
Obtain the obstacle set $\mathcal{X}_{obstacle}$ from the train set $\mathcal{X}_{train}$.
%*Training stage*
Initialize the CNN network parameters Θ.
for each mini-batch $\mathcal{B} \subset \mathcal{X}_{train}$ **do**
  Create set $\mathcal{B}' = \varnothing$
  Random select obstacle $[\mathbf{o}_1, ..., \mathbf{o}_k] \in \mathcal{X}_{obstacle}$.
  for each $\mathbf{x}_i \in \mathcal{B}$ **do**
    Generate augmented sample $[\mathbf{x}_{i,1}, ..., \mathbf{x}_{i,k}]$ by $\mathbf{x}_i$ and $[\mathbf{o}_1, ..., \mathbf{o}_k]$.
    Adding $[\mathbf{x}_i; \mathbf{x}_{i,1}, ..., \mathbf{x}_{i,k}]$ to set $\mathcal{B}'$.
  end for
  Extract ID relevant feature $\mathbf{f}_i$ and irrelevant feature $\bar{\mathbf{f}}_i$ of each $\mathbf{x}_i \in \mathcal{B}'$ using DRL-Net.
  Calculate $\mathcal{L}_{ce}, \mathcal{L}_o, \mathcal{L}_{tri}, \mathcal{L}_{rtri}$ by Eqs.(1, 2, 3, 4).
  Optimize CNN parameters Θ according to Eq.(5).
end for
%*Inference stage*
for each $\mathbf{x}_q \in \mathcal{X}_{query}$, $\mathbf{x}_g \in \mathcal{X}_{gallery}$ **do**
  Extract $\mathbf{f}_q, \mathbf{f}_g$ of $\mathbf{x}_q, \mathbf{x}_g$ respectively using DRL-Net.
  Calculate $\mathcal{D}(\mathbf{f}_q, \mathbf{f}_g)$ by cosine distance metric.
end for
return $\mathcal{D}$

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

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 [37] 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** [38] consists of 32,668 images of 1,501 identities captured by 6 camera views. Following the standard setting [38], the whole dataset is divided into a training set containing 12,936 images of 751 identities and a testing set containing 19,732 images of 750 identities. **DukeMTMC** [37] contains of 36,411 images of 1,812 persons from 8 cameras. 16,522 images of 702 persons are randomly selected from the dataset as the training set, and the remaining images are divided into the testing set containing 2,228 query images and 17,661 gallery images. The setting is same to [37]. **MSMT17** contains 126,441 images of 4,101 IDs captured from a 15-camera network. The training set has 32,621 images of 1,041 identities, and the testing set has 93,820 images of 3,060 identities. During inference, 11,659 images are randomly selected as query images and the other 82,161 images are used as gallery images from the testing set [39].

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 matching results.

B. Implementation Details

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 [40] 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 [41] as the convolution neural backbone network. Following the setting of most ReID methods [42] , 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 $2048 \times 16 \times 8$. The hidden dimension d is set to 256. The transformer layers are same with DETR and initialized with Xavier init [43]. 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 probe and gallery images.

Optimization. The CNN backbone network is pertrained over ImageNet [44]. 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.

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 III, 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 comparision results over dataset Occluded-DukeMTMC are shown in Table I. There are three mainstream types of occluded re-ID methods are compared: holistic re-ID methods without designing modules for occlusions (DIM [45], Part Aligned [46], HACNN [47], Adver Occluded [48] and PCB [23]), Occluded re-ID methods with external cues (Part Bilinear [49], FD-GAN [50], PGFA [12] and HONet [17]), and Occluded re-ID methods based on part-to-part matching (DSR [19], SFR [51] and MoS [52]). Our DRL-Net method achieves 65.0% Rank-1 accuracy,

TABLE I: Comparison with state-of-the-art methods of occluded re-ID on Occluded-DukeMTMC. It shows that DRL-Net is superior to all three types of methods including the methods designed for holistic re-ID in the 1st group, utilizing external cues in the 2nd group and adopting part-to-part matching strategy in the 3rd group.

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	-	-	32.2
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
MoS (AAAI 21)	61.0	74.4	79.1	49.2
DRL-Net $(Ours)$	**65.0**	**79.3**	**83.6**	**50.8**

TABLE II: Ablation study over Occluded-DukeMTMC. T, OSA and CFL denotes proposed transformer architecture, occluded sample augmentation and contrast feature learning respectively.

Methods	Rank-1	Rank-5	Rank-10	mAP
Baseline	51.0	66.3	71.7	43.8
Baseline+T	59.6	74.4	79.5	49.0
Baseline+T+OSA	60.5	74.7	80.9	48.2
Baseline+T+OSA+CFL	65.0	79.3	83.6	50.8

79.3% Rank-5 accuracy, 83.6% Rank-10 accuracy, and 50.8% mAP, which outperforming all types of methods by a large margin.

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 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.

Results on Market-1501 and DukeMTMC. The comparision results over holistic re-ID datasets including Market-1501 and Duke-MTMC are shown in Table III. Three types of holistic re-ID methods are considered for comparison: methods based on global features (IANet [53], MVPM [54], DMML [55], SFT [56], VCFL [57], Circle [58], MoS [52]), 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 [59]) and human pose based (Pose-transfer [60], PSE [61], PGFA [12] and HONet [17]). Though DRL-Net is not proposed for the holistic re-ID task, it performs comparable results with all types of holistic re-ID methods, showing the robustness of our proposed methods.

Results on MSMT17. Since MSMT17 is released recently, hence there are only a few methods that report on this dataset, including MVMP [54], SFT [56], DG-Net [62], IANet [53], Circle [58] and Circle + MGN [58]. Table IV shows the comparison results. DRL-Net achieves outstanding performance in all evaluation metrics.

TABLE III: 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.

Methods	Market-1501		DukeMTMC	
	Rank-1	mAP	Rank-1	mAP
IANet (CVPR 19)	94.4	83.1	87.1	73.4
MVPM (ICCV 19)	91.4	80.5	83.4	70.0
DMML (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	-	-
PCB(ECCV 18)	92.3	77.4	81.8	66.1
PCB+RPP (ECCV 18)	93.8	81.6	83.3	69.2
AlignedReID(ArXiv18)	91.8	79.3	-	-
DSR (CVPR 18)	83.6	64.3	-	-
VPM (CVPR 19)	93.0	80.8	83.6	72.6
SPReID (CVPR 18)	92.5	81.3	-	-
MGCAM (CVPR 18)	83.8	74.3	46.7	46.0
Pose-transfer (CVPR18)	87.7	68.9	30.1	28.2
PSE (CVPR 18)	87.7	69.0	27.3	30.2
PGFA (ICCV 19)	91.2	76.8	82.6	65.5
AANet (CVPR 19)	93.9	82.5	86.4	72.6
HONet (CVPR 20)	94.2	84.9	86.9	75.6
DRL-Net *(Ours)*	**94.7**	**86.9**	**88.1**	**76.6**

TABLE IV: Comparison over holistic re-ID dataset MSMT17.

Methods	Rank-1	Rank-5	Rank-10	mAP
MVPM (ICCV 19)	71.3	84.7	-	46.3
SFT (ICCV 19)	73.6	85.6	-	47.6
DG-Net (CVPR 19)	77.2	87.4	90.5	52.3
IANet (CVPR 19)	75.5	85.5	88.7	46.8
Circle (CVPR 20)	76.3	-	-	50.2
Circle + MGN (CVPR 20)	76.9	-	-	52.1
DRL-Net $(Ours)$	**78.4**	**88.2**	**91.3**	**55.3**

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. Table II shows experimental results.

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 II, 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.

TABLE V: The comparison and analysis for the number N_l of transformer layers over Occluded-DukeMTMC. The number of layers for encoder and decoder are the same.

N_l	Rank-1	Rank-5	Rank-10	mAP
1	55.3	71.8	78.0	47.0
2	57.2	72.4	78.2	47.0
3	55.9	71.7	77.4	46.0
4	55.3	71.3	77.6	45.2
5	55.7	73.2	78.8	44.9
6	53.9	70.1	75.7	42.9

TABLE VI: Parameter analysis of the number N_q of object queries over Occluded-DukeMTMC. The results shows DRL-Net is robust to different N_q.

N_q	Rank-1	Rank-5	Rank-10	mAP
2	53.5	69.2	75.3	43.6
5	55.3	71.9	77.8	46.4
9	57.1	72.4	78.2	47.3
13	57.2	73.9	79.0	47.3
17	57.7	75.2	79.7	47.2

Effectiveness of the proposed OSA. We evaluate the occluded sample augmentation as described in Section III-B1. 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 II, 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-B1. 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 shown in Table II, the incorporation of contrast feature learning significantly improves the person re-ID performance beyond *baseline+T+OSA*. The *baseline+T+OSA+CFL* achieves a rank-1 accuracy of 65.0% and an mAP of 50.8% Which outperforms the corresponding *baseline+T+OSA* by 4.5% and 2.6%, respectively. The effectiveness of the contrast feature learning 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 14.0% in Rank-1 accuracy and 7.0% 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.

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 V. 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.

Fig. 3: Illustration of the proposed DRL-Net: For each occluded query person image on the left, the two rows on the right show the ten top-matching images as returned by the *baseline* model (in the first row) and the proposed DRL-Net (in the second row). The green and red boxes highlight positive and negative matching, respectively.

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

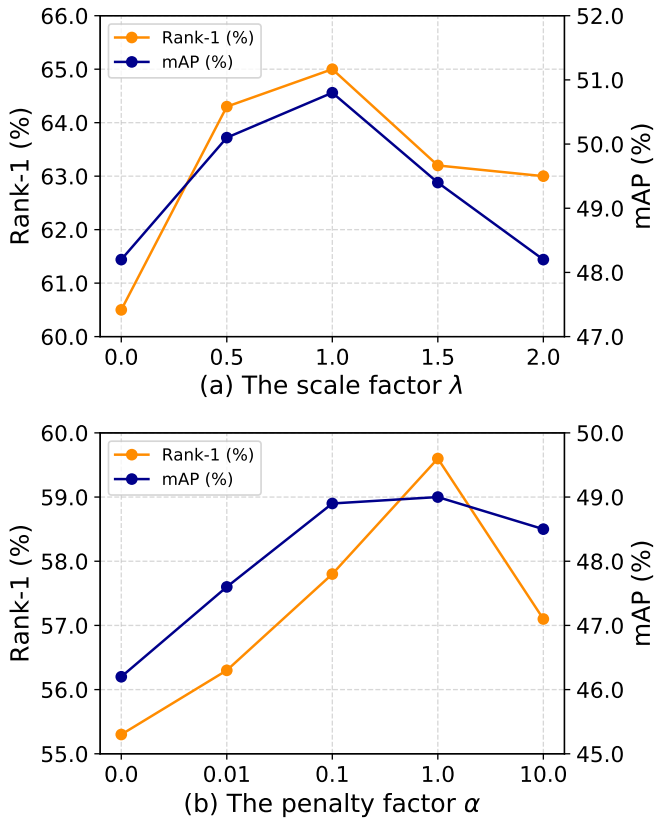

Fig. 4: Parameter analysis for the scale factor λ and the penalty factor α.

suitable N_q. As detailed in Table VI, 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 set N_q to 9 in the final version.

The Robustness of Parameters λ and α. 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 Eq. 5 is the balancing weight of contrast feature learning (CFL). With λ increasing, the Rank-1/mAP is improved by 4.5%/2.6% (λ = 1.0), which means the CFL module now is beneficial for learning better occlusion robust re-ID features. Continuing to increase λ, 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 α on *baseline+T* model as shown in Fig 4 (b). The penalty factor α in Eq. 2 will affect correlation among object queries in decoder. Experiments show that *baseline+T* performs best when α = 1.0. Using a smaller α will suppress the value of $\mathcal{L}_o$, lower the decorrelation ability for object queries. On the other hand, α 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.

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.

F. Visualization

Qualitative Results. We demonstrate how DRL-Net overcomes the occlusion constraint by providing several samples of person image ranking and Fig 3 shows experimental results. For each occluded query person image on the left, the two rows of images on the right show the 10 top-matching images that are produced by the *baseline* and our proposed DRL-Net, respectively. We can observe that DRL-Net can overcome the occlusions and identify images of the same pedestrian correctly (highlighted by green-color boxes). As a comparison, the *baseline* network is very sensitive to occlusions obviously and returns a large amount of false-matching person images (highlighted by red-color boxes).

Visualizing for object query. We visualize decoder cross-attentions for the ID-irrelevant object query (the last object query on decoder) using the heat map. The redder the heat map, the higher the attention scores. As Fig. 5 shows, the ID-irrelevant object query can automatically localize occlusion and noise areas without explicit supervision, though obstacles various in different person images. This nice property is largely attributed to the transformer architecture and the contrast feature learning that guides the network to disentangle the representation of different semantic components and encourage the separation of occlusions feature and discriminative ID-relevant features.

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," *arXiv preprint arXiv:1610.02984*, 2016.

[2] S. Gong, M. Cristani, C. C. Loy, and T. M. Hospedales, "The re-identification challenge," in *Person re-identification*. 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 Transactions on Multimedia (TMM)*, vol. 20, no. 3, pp. 593–604, 2017.

[4] Z. Wang, J. Jiang, Y. Yu, and S. Satoh, "Incremental re-identification by cross-direction and cross-ranking adaption," *IEEE Transactions on Multimedia (TMM)*, vol. 21, no. 9, pp. 2376–2386, 2019.

[5] H. Luo, W. Jiang, Y. Gu, F. Liu, X. Liao, S. Lai, and J. Gu, "A strong baseline and batch normalization neck for deep person re-identification," *IEEE Transactions on Multimedia (TMM)*, vol. 22, no. 10, pp. 2597–2609, 2019.

[6] M. Cao, C. Chen, H. Dou, X. Hu, S. Peng, and A. Kuijper, "Progressive bilateral-context driven model for post-processing person re-identification," *IEEE Transactions on Multimedia (TMM)*, vol. 23, pp. 1239–1251, 2020.

[7] Z. Wang, R. Hu, C. Liang, Y. Yu, J. Jiang, M. Ye, J. Chen, and Q. Leng, "Zero-shot person re-identification via cross-view consistency," *IEEE Transactions on Multimedia (TMM)*, vol. 18, no. 2, pp. 260–272, 2015.

[8] Z. Zeng, Z. Wang, Z. Wang, Y. Zheng, Y.-Y. Chuang, and S. Satoh, "Illumination-adaptive person re-identification," *IEEE Transactions on Multimedia (TMM)*, vol. 22, no. 12, pp. 3064–3074, 2020.

[9] Y. Wang, R. Hu, C. Liang, C. Zhang, and Q. Leng, "Camera compensation using a feature projection matrix for person reidentification," *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, vol. 24, no. 8, pp. 1350–1361, 2014.

[10] L. Wei, S. Zhang, H. Yao, W. Gao, and Q. Tian, "Glad: Global–local-alignment descriptor for scalable person re-identification," *IEEE Transactions on Multimedia*, vol. 21, no. 4, pp. 986–999, 2018.

[11] J. Zhuo, Z. Chen, J. Lai, and G. Wang, "Occluded person re-identification," in *Proceedings of the IEEE conference on Multimedia and Expo (ICME)*, 2018, pp. 1–6.

[12] J. Miao, Y. Wu, P. Liu, Y. Ding, and Y. Yang, "Pose-guided feature alignment for occluded person re-identification," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019, pp. 542–551.

[13] C. Song, Y. Huang, W. Ouyang, and L. Wang, "Mask-guided contrastive attention model for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 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 *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 1062–1071.

[16] S. Gao, J. Wang, H. Lu, and Z. Liu, "Pose-guided visible part matching for occluded person reid," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 11 744–11 752.

[17] G. Wang, S. Yang, H. Liu, Z. Wang, Y. Yang, S. Wang, G. Yu, E. Zhou, and J. Sun, "High-order information matters: Learning relation and topology for occluded person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 6449–6458.

[18] W.-S. Zheng, X. Li, T. Xiang, S. Liao, J. Lai, and S. Gong, "Partial person re-identification," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7073–7082.

[20] Y. Sun, Q. Xu, Y. Li, C. Zhang, Y. Li, S. Wang, and J. Sun, "Perceive where to focus: Learning visibility-aware part-level features for partial person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 393–402.

[21] H. Luo, W. Jiang, X. Fan, and C. Zhang, "Stnreid: Deep convolutional networks with pairwise spatial transformer networks for partial person re-identification," *IEEE Transactions on Multimedia (TMM)*, vol. 22, no. 11, pp. 2905–2913, 2020.

[22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems (NIPS)*, 2017, pp. 5998–6008.

[23] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, "Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 480–496.

[24] X. Zhang, H. Luo, X. Fan, W. Xiang, Y. Sun, Q. Xiao, W. Jiang, C. Zhang, and J. Sun, "Alignedreid: Surpassing human-level performance in person re-identification," *ArXiv*, vol. abs/1711.08184, 2017.

[25] K. Zhu, H. Guo, Z. Liu, M. Tang, and J. Wang, "Identity-guided human semantic parsing for person re-identification," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020, pp. 346–363.

[26] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2019.

[27] A. van den Oord, Y. Li, I. Babuschkin, K. Simonyan, O. Vinyals, K. Kavukcuoglu, G. van den Driessche, E. Lockhart, L. Cobo, F. Stimberg, N. Casagrande, D. Grewe, S. Noury, S. Dieleman, E. Elsen, N. Kalchbrenner, H. Zen, A. Graves, H. King, T. Walters, D. Belov, and D. Hassabis, "Parallel WaveNet: Fast high-fidelity speech synthesis," in *Proceedings of the 35th International Conference on Machine Learning (ICML)*, 2018, pp. 3918–3926.

[28] J. Gu, J. Bradbury, C. Xiong, V. O. Li, and R. Socher, "Non-autoregressive neural machine translation," in *Proceedings of International Conference on Learning Representations (ICLR)*, 2018.

[29] M. Ghazvininejad, O. Levy, Y. Liu, and L. Zettlemoyer, "Mask-predict: Parallel decoding of conditional masked language models," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019.

[30] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020, pp. 213–229.

[31] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proceedings of International Conference on Learning Representations (ICLR)*, 2020.

[32] H. Chen, Y. Wang, T. Guo, C. Xu, Y. Deng, Z. Liu, S. Ma, C. Xu, C. Xu, and W. Gao, "Pre-trained image processing transformer," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 12 299–12 310.

[33] E. Xie, W. Wang, W. Wang, P. Sun, H. Xu, D. Liang, and P. Luo, "Segmenting transparent object in the wild with transformer," vol. abs/2101.08461, 2021.

[34] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. Torr *et al.*, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 6881–6890.

[35] L. Huang, J. Tan, J. Liu, and J. Yuan, "Hand-transformer: Non-autoregressive structured modeling for 3d hand pose estimation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020, pp. 17–33.

[36] L. Huang, J. Tan, J. Meng, J. Liu, and J. Yuan, "Hot-net: Non-autoregressive transformer for 3d hand-object pose estimation," *Proceedings of the 28th ACM International Conference on Multimedia (ACM MM)*, pp. 3136–3145, 2020.

[37] Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled samples generated by gan improve the person re-identification baseline in vitro," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 3754–3762.

[38] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1116–1124.

[39] L. Wei, S. Zhang, W. Gao, and Q. Tian, "Person transfer gan to bridge domain gap for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 79–88.

[40] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2020, pp. 13 001–13 008.

[41] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[42] H. Luo, Y. Gu, X. Liao, S. Lai, and W. Jiang, "Bag of tricks and a strong baseline for deep person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPR Workshops)*, 2019, pp. 1487–1495.
[43] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the thirteenth international conference on Artificial Intelligence and Statistics (AISTATS)*, 2010, pp. 249–256.
[44] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems (NIPS)*, pp. 1097–1105, 2012.
[45] Q. Yu, X. Chang, Y.-Z. Song, T. Xiang, and T. M. Hospedales, "The devil is in the middle: Exploiting mid-level representations for cross-domain instance matching," *ArXiv*, vol. abs/1711.08106, 2017.
[46] L. Zhao, X. Li, Y. Zhuang, and J. Wang, "Deeply-learned part-aligned representations for person re-identification," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 3219–3228.
[47] W. Li, X. Zhu, and S. Gong, "Harmonious attention network for person re-identification," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 2285–2294.
[48] H. Huang, D. Li, Z. Zhang, X. Chen, and K. Huang, "Adversarially occluded samples for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 5098–5107.
[49] Y. Suh, J. Wang, S. Tang, T. Mei, and K. Mu Lee, "Part-aligned bilinear representations for person re-identification," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 418–437.
[50] Y. Ge, Z. Li, H. Zhao, G. Yin, S. Yi, X. Wang *et al.*, "Fd-gan: Pose-guided feature distilling gan for robust person re-identification," in *Advances in Neural Information Processing Systems (NIPS)*, 2018, pp. 1230–1241.
[51] L. He, Z. Sun, Y. Zhu, and Y. Wang, "Recognizing partial biometric patterns," *ArXiv*, vol. abs/1810.07399, 2018.
[52] M. Jia, X. Cheng, Y. Zhai, S. Lu, S. Ma, Y. Tian, and J. Zhang, "Matching on sets: Conquer occluded person re-identification without alignment," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2021, pp. 1673–1681.
[53] R. Hou, B. Ma, H. Chang, X. Gu, S. Shan, and X. Chen, "Interaction-and-aggregation network for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 9317–9326.
[54] H. Sun, Z. Chen, S. Yan, and L. Xu, "Mvp matching: A maximum-value perfect matching for mining hard samples, with application to person re-identification," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019, pp. 6737–6747.
[55] G. Chen, T. Zhang, J. Lu, and J. Zhou, "Deep meta metric learning," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019, pp. 9547–9556.
[56] C. Luo, Y. Chen, N. Wang, and Z. Zhang, "Spectral feature transformation for person re-identification," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019, pp. 4976–4985.
[57] F. Liu and L. Zhang, "View confusion feature learning for person re-identification," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019, pp. 6639–6648.
[58] Y. Sun, C. Cheng, Y. Zhang, C. Zhang, L. Zheng, Z. Wang, and Y. Wei, "Circle loss: A unified perspective of pair similarity optimization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 6398–6407.
[59] C.-P. Tay, S. Roy, and K.-H. Yap, "Aanet: Attribute attention network for person re-identifications," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 7134–7143.
[60] J. Liu, B. Ni, Y. Yan, P. Zhou, S. Cheng, and J. Hu, "Pose transferrable person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 4099–4108.
[61] M. S. Sarfraz, A. Schumann, A. Eberle, and R. Stiefelhagen, "A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 420–429.
[62] Z. Zheng, X. Yang, Z. Yu, L. Zheng, Y. Yang, and J. Kautz, "Joint discriminative and generative learning for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 2138–2147.