OCCLUDED PERSON RE-IDENTIFICATION VIA RELATIONAL ADAPTIVE FEATURE CORRECTION LEARNING

Minjung Kim, MyeongAh Cho, Heansung Lee, Suhwan Cho, Sangyoun Lee

School of Electrical and Electronic Engineering, Yonsei University

ABSTRACT

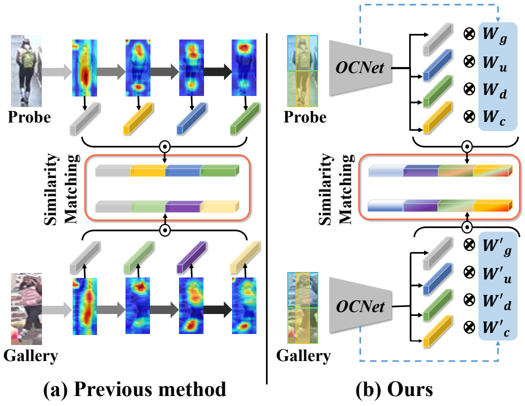
Occluded person re-identification (Re-ID) in images captured by multiple cameras is challenging because the target person is occluded by pedestrians or objects, especially in crowded scenes. In addition to the processes performed during holistic person Re-ID, occluded person Re-ID involves the removal of obstacles and the detection of partially visible body parts. Most existing methods utilize the off-the-shelf pose or parsing networks as pseudo labels, which are prone to error. To address these issues, we propose a novel Occlusion Correction Network (OCNet) that corrects features through relational-weight learning and obtains diverse and representative features without using external networks. In addition, we present a simple concept of a center feature in order to provide an intuitive solution to pedestrian occlusion scenarios. Furthermore, we suggest the idea of Separation Loss (SL) for focusing on different parts between global features and part features. We conduct extensive experiments on five challenging benchmark datasets for occluded and holistic Re-ID tasks to demonstrate that our method achieves superior performance to state-of-the-art methods especially on occluded scene.

Index Terms— Occluded Person Re-Identification, Relation Network, Person Re-Identification, Deep Learning

1. INTRODUCTION

Person re-identification (Re-ID) involves identifying people appearing in the images captured by multiple cameras with non-overlapping domains. Most of methods utilize a representation by extracting global features from complete pedestrian images. However, these approaches have limitations in addressing the occluded Re-ID, which frequently occurs in crowded and complex scenes featuring multiple obstacles.

Occlusion is classified into object occlusion and pedestrian-Interference (PI) [1] based on the type of object that covers the target. In object occlusion cases, the global feature extracted by the network may be mixed with unnecessary information.



(a) Previous method **(b) Ours**

Fig. 1: Challenge related to feature misalignment and our proposed OCNet. (a) Previous approaches suffered from feature misalignment problem when calculating the feature similarity because features are extracted in the order of high activation. (b) OCNet aligns various features in advance and then corrects the features with weight features W obtained from relational information.

In order to obtain only the target features, most methods [2, 3] use attention maps or pose information generated by other additional networks such as pose estimation and human parsing networks. However, these methods are highly dependent on the performance of the external networks and are hard to train well with an error-prone pseudo label. Without such external help, it is impossible to flexibly deal with misalignment problems. The misalignment problem occurs when features are extracted in the order of important parts from the input image and simply concatenated. For example, in the Fig. 1 (a), part features are extracted and concatenated in body-feet-upper body-head order and body-head-upper body-feet order for probe and gallery images; This leads to misalignment of the final representation features, lowering similarity and unmatching IDs. Another issue is PI, which occurs when two or more people appear in the bounding box because of the limitations of detector performance in crowded scenes. As the network recognizes multiple pedestrians as the foreground, it is difficult to remove the non-target pedestrians. This phenomenon leads to mismatching between people and IDs. In existing methods [1, 4], external networks are used or an at-

This work was supported by the Institute of Information communications Technology Planning Evaluation(IITP) grant funded by the Korean government(MSIT) (No. 2021-0-00172, The development of human Re-identification and masked face recognition based on CCTV camera)

978-1-6654-0540-9/22/$31.00 ©2022 IEEE

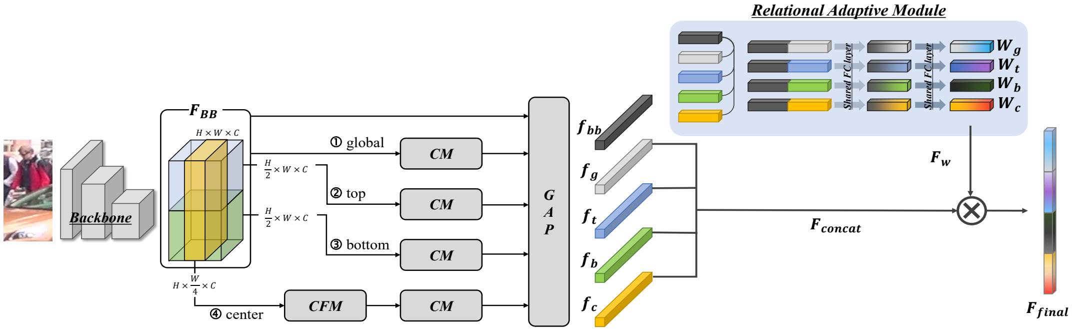


Fig. 2: Overall framework of the proposed approach. OCNet consists of Converter Module (CM), Center-Focus module (CFM), and Relation Adaptive Module (RAM). The backbone feature $\mathbf{F}_{BB}$ go through CM to extract global features $\mathbf{f}_g$, top features $\mathbf{f}_t$, bottom features $\mathbf{f}_b$, and center features $\mathbf{f}_c$. After passing through RAM, it becomes a relational-weight $\mathbf{F}_w$ with the same size as $\mathbf{F}_{concat}$. Multiplying these two results in a final representation feature that is well-aligned and robust to occlusion.

tention map is obtained by performing additional operations upon query and gallery image to find the target.

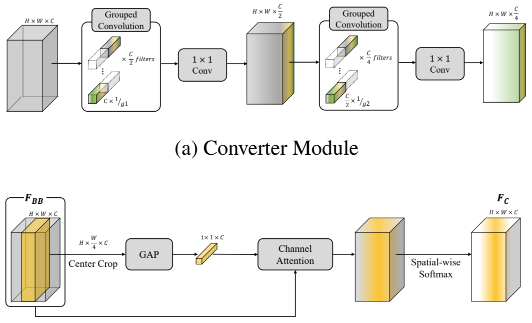
To address the aforementioned issues, we propose Occlusion Correction Network (OCNet), a novel network that extracts global and various part features while performing feature correction in order to be robust to occlusion. Unlike the previous method [5] of simply concatenating the part features, which causes feature misalignment, our model predicts the weights through relational information between various features and corrects irrelevant information to create the final representation. Through relational-weight learning, the weights of the occluded part are adaptively determined and aligned with feature correction. Thus, our method extracts various features from the visible part of the given images even if occluded; it then characterizes the relationship between them to create a final representation feature that is robust to occlusion and contains a discriminative representation.

In this paper, our main contributions are as follows: (1) We design an **Occlusion Correction Network** that corrects the final representation through relational-weight learning to determine whether a part feature is crucial or not, so that a more diverse and representative feature is obtained which avoids occluded parts. (2) To provide a robust solution to pedestrian occlusion, we propose the essential concept of a **center feature** for the first time to provide a robust solution to pedestrian occlusion. (3) We suggest **Separation Loss** for focusing on different information between global features and part features. (4) Our network achieves **superior performance** not only in occluded dataset but also in holistic datasets, without relying on additionally trained networks or external annotations.

2. PROPOSED METHOD

2.1. Occlusion Correction Network

A practical solution for occluded Re-ID is to extract as many features as possible from the non-occluded regions of the given images. We extract four features ($\mathbf{f}_g$, $\mathbf{f}_t$, $\mathbf{f}_b$, $\mathbf{f}_c$) in various ways. Given that these features are generated from $\mathbf{F}_{BB}$, which is activated in the foreground, they are not substantially affected by background cluttering. $\mathbf{f}_g$ is obtained through a CM so that the network finds more important features in the existing $\mathbf{F}_{BB}$. $\mathbf{f}_t$ and $\mathbf{f}_b$ are created by cropping the feature map extracted from the last layer of the backbone into two parts, top and bottom. OCNet learns through SL so that $\mathbf{f}_t$ and $\mathbf{f}_b$ can learn meaningful part-level features. $\mathbf{f}_c$ generated by CFM is robust in PI because it focuses on the target. As shown in Fig. 2, we concatenate all these features ($\mathbf{f}_g$, $\mathbf{f}_t$, $\mathbf{f}_b$, $\mathbf{f}_c$) to form $\mathbf{F}_{concat}$ and multiply the weight vector $\mathbf{F}_w$ predicted by RAM. Therefore, we obtain the final representation feature $\mathbf{F}_{final}$ in which the feature misalignment is solved.



(a) Converter Module

(b) Center-Focus Module

Fig. 3: Overview of our (a) Converter Module (CM) (b) Center-Focus Module (CFM).

Converter Module. For the part features of the pedestrian to have differently activated for each channel, a structure that divides and calculates a group of channels is needed. We adopt grouped convolution, which has the advantage of learning a channel with high correlation for each group. As shown in Fig. 3 (a). CM not only reduces channels based on the high response of channels but also provides parameters that each feature learns with the corresponding objective function.

Table 1: Comparison with state-of-the-art methods on occluded Re-ID dataset [3] and holistic Re-ID datasets [6, 7, 8]. Methods are divided into 4 groups: holistic-reid-targeted, attention-based, occlusion-targeted with external-cues-based and occlusion-targeted with no external-cues-based. The 2nd highest performance is underlined. ”*” denotes different backbone.

Method	Occluded DukeMTMC		Market1501		DukeMTMC-reID		CuHK03-L		CuHK03-D	
	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
PCB [5] (ECCV 2018)	42.6	33.7	92.3	77.4	81.8	66.1	-	-	63.7	57.5
BoT [9] (CVPRW 2019)	47.7	41.3	94.1	85.7	86.4	76.4	-	-	-	-
CASN [10] (CVPR 2019)	-	-	94.4	82.8	87.7	73.7	73.7	68	71.5	64.4
CAMA [11] (CVPR 2019)	-	-	94.7	84.5	85.8	72.9	70.1	66.5	66.6	64.2
PGFA [3] (ICCV 2019)	51.4	37.3	91.2	76.8	82.6	65.5	-	-	-	-
HoReID [12] (CVPR 2020)	55.1	43.8	94.2	84.9	86.9	75.6	-	-	-	-
OCNet (ours)	59.9	49.7	94.9	87.2	87.8	77.2	77.9	74.9	76.7	72.4
ISP* [13] (ECCV 2020)	62.8	52.3	95.3	88.6	88.7	78.9	76.5	74.1	75.2	71.4
MoS* [14] (AAAI 2021)	**66.6**	**55.1**	**95.4**	89	**90.6**	80.2	-	-	-	-
OCNet + ibn *(ours)	64.3	54.4	95	**89.3**	90.5	**80.2**	**82**	**78.6**	**78.9**	**76.2**

Center-Focus Module. In the PI problem, it is unclear which ID should be assigned as a label when the target is largely obscured by other people. In this case, using the detector mechanism, i.e., the natural assumption that the target is located in the middle of the bounding box, the feature focusing on the corresponding center part helps to find the target. CFM extracts features using the attention technique that weights the feature located in the middle of the bounding box. Unlike SENet [15], CFM crops the part corresponding to the center in the spatial domain of $\mathbf{F}_{BB}$ and the softmax is taken to form a spatial-level probability map as shown in Fig. 3 (b). Therefore, we extract a feature focusing on the center by multiplying and adding with $\mathbf{F}_{BB}$.

2.2. Relational Adaptive Module

Most Re-ID methods [5, 16] concatenate various part features to make one final representation feature and use it for learning. However, it is not known what information each feature has, so a feature misalignment problem can occur when calculating the final feature distance. To solve feature misalignment, a correction process must be performed to estimate the contribution of each feature and reflect it to generate the final representation feature. We propose a RAM, which is different from relation network [17, 18], for estimating weights with relational information between features. (It is covered in 3.2.) RAM consists of two fully connected shared layers that define relationships between vectors and handle occluded or insignificant features by weights. The $\mathbf{F}_w$ generated in RAM is multiplied by $\mathbf{F}_{concat}$ to finally correct the feature.

2.3. Separation Loss Function

It is crucial to extract part-level features that are as diverse as possible from the parts representing the identity. We introduce a Separation Loss (SL) that prevents excessive focus on a specific region in an image with a single ID. SL calculates cosine similarity between ($\mathbf{f}_g$, $\mathbf{f}_t$) and ($\mathbf{f}_g$, $\mathbf{f}_b$) pairs as described in Eq. 1.

$$L_{SL} = (1 + \frac{\mathbf{f}_g \cdot \mathbf{f}_t}{\|\mathbf{f}_g\|_2 \, \|\mathbf{f}_t\|_2}) + (1 + \frac{\mathbf{f}_g \cdot \mathbf{f}_b}{\|\mathbf{f}_g\|_2 \, \|\mathbf{f}_b\|_2}) \tag{1}$$

$\mathbf{f}_t$ and $\mathbf{f}_b$ are already uniformly horizontal partitions of $\mathbf{F}_{BB}$. That is, they already have different information because they are physically separated. Therefore, SL learns to have different information because the value increases as $\mathbf{f}_g$, which has global information, shares information with $\mathbf{f}_t$ and $\mathbf{f}_b$.

2.4. Training and Testing Phases

In the training phase, OCNet employs three kinds of losses, denoted as L_{ID}, L_{Tri}, and L_{SL}. As describes in Eq. 2, L_{ID} and L_{Tri} is the sum of L_{ce} of $\mathbf{F}_{final}$, $\mathbf{F}_{BB}$, $\mathbf{f}_g$, and $\mathbf{f}_c$ where L_{ce} denotes cross-entropy with label smooth, and L_t denotes triplet loss with margin, respectively. The overall objective function is defined as Eq. 3.

$$\begin{aligned} L_{ID/Tri} = \lambda_1 L_{ce/t}(\mathbf{F}_{final}) + \lambda_2 L_{ce/t}(\mathbf{F}_{BB}) + \\ \lambda_3 L_{ce/t}(\mathbf{f}_g) + \lambda_4 L_{ce/t}(\mathbf{f}_c) \end{aligned} \tag{2}$$

$$L_{total} = L_{ID} + L_{Tri} + \gamma L_{SL} \tag{3}$$

In the testing phase, we calculate l_2 distances of $\mathbf{F}_{final}$ and $\mathbf{F}_{BB}$. Then, as in Eq. 4, we consider both the distances of $\mathbf{F}_{final}$ containing relational information and the distances of $\mathbf{F}_{BB}$ containing general information from backbone.

$$D = \left\| \mathbf{F}^q_{final} - \mathbf{F}^g_{final} \right\|_2 + \alpha \left\| \mathbf{F}^q_{BB} - \mathbf{F}^g_{BB} \right\|_2 \tag{4}$$

3. EXPERIMENTAL RESULTS

Implementation Details: We adopt ResNet-50 as a backbone, pretrained using ImageNet. We refer to the structure in which a fully connected layer follows the batch normalization layer in and apply effective training strategies to use it as a baseline model [9]. $\lambda 1$, $\lambda 2$, $\lambda 3$, $\lambda 4$ is set to 0.8, 0.5, 0.25, 0.25 respectively, depending on the importance of the feature. γ and α are set to 1 in our experiments.

Table 2: Analysis of Relation Network (RN), concatenated features, and OCNet modules (CFM, SL, RAM).

Index		1	2	3	4	5	6	7	8	9
RN						✓	✓			
Concat				✓	✓	✓	✓	✓	✓	✓
OCNet	CFM		✓			✓	✓	✓		✓
	SL				✓		✓		✓	✓
	RAM							✓	✓	✓
Rank-1		47.7	49.2	55.2	58.6	52.7	53	52.9	58.6	**59.9**
mAP		41.3	42	45.9	49.1	44.2	43.2	45.1	49.5	**49.7**

3.1. Comparison with State-of-the-art Methods

To show that our proposed method is effective not only in occlusion scenarios but also in general cases, we conduct experiments on two scenarios which is summarized in Table 1.

Results on Occluded Re-ID dataset. For a fair comparison, we compared with the 4th group using the ResNet-50 with ibn layer, which improves the generalization capacity of the model without increasing the model complexity. 1st group suffers when crucial information is occluded and 3rd group has no clear solution for PI, which leads to ID mismatching. OCNet has a lower performance than MoS but has the 2nd performance without additional operation in the inference.

Results on Holistic Re-ID datasets. Even though our OCNet is occluded-targeted method, it outperforms the most recent state-of-the-art method on holistic datasets. In particular, in CUHK03-D, the bounding box is not tight, so multiple people often appear in the image. Our method exceeds state-of-the-art methods in the dataset corresponding to the problem we are trying to solve.

3.2. Discussion

We conduct extensive ablation studies on Occluded-Duke MTMC. It is summarized in Table 2.

OCNet. When comparing Index-1 and Index-3, extracting various features enables us to discriminatively express identity. Index-9, where all of our proposed modules are applied, shows a significant difference in performance.

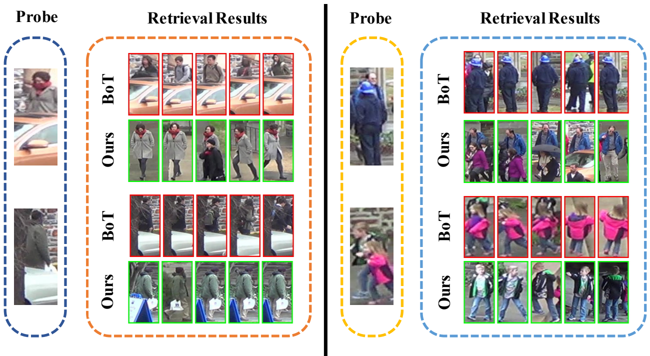
CFM & SL. Index-2 extracts features focused on the center by CFM. Index-8 removes CFM from OCNet. When comparing Index-1 and Index-2, extracting the feature through CFM helps to find the target. From index-7 and index-9, it indicates that SL enables the extraction of various features.

RAM. Index-5 learns using vectors containing relational information by applying the existing Relation Network (RN)[17]. A comparison of index-6 and index-9 shows the difference between our RAM and the RN. If the relation vectors are used in occlusion scenarios, unnecessary information is intertwined. By contrast, RAM predicts weights based on relational information and corrects features to generate discriminative features that are robust against occlusion.

Influence of the Size of Center. From Table 3, the best performance is when W is 2. This is because, contrary to the intention of CFM, as the size of the feature map to focus increases, information about non-target objects gets mixed in.

Table 3: Comparison with varing settings of hyperparameters on Occluded-DukeMTMC.

Hyper-	Size of Center				Num of Parts				γ			
param.	2	4	6	8	1	2	3	4	0	1	2	5
Rank-1	**59.9**	58.1	59.2	59	57.8	**59.9**	59.4	59.5	52.9	**59.9**	57.7	55.6
mAP	**49.7**	49.3	49.4	49.6	48.2	**49.7**	49.2	49.9	45.1	**49.7**	48.5	47



(a) Object Occlusion

(b) Pedestrian-Interference

Fig. 4: Ranking results of BoT and OCNet. The red and green bounding indicate the error and correct result, respectively.

Influence of the Number of Parts. The more parts there are, the more modules need to be added, and the size of the feature also increases. Considering this, we set it to two parts.

Analysis of Hyperparameters of SL. From Table 3, when γ is 1, the best performance is achieved. SL must be smaller than the other losses to avoid interfering with the identification objective.

3.3. Visualization

As shown in Fig. 4 (a), BoT [9] has a tendency to recognize the obstacle as a part of the feature, so images where the obstacle and the target coexist, are retrieved. However, OCNet which corrects the occluded part feature successfully retrieval by concentrating only on the features corresponding to the target. The Fig. 4 (b) shows pedestrian-interference samples, where the non-target occludes the target significantly. BoT is biased towards non-targets that are more visible than the target. On the contrary, OCNet successfully retrievals the target.

4. CONCLUSION

In this paper, we propose an Occlusion Correction Network (OCNet) that corrects features through relational-weight learning and obtains diverse and representative features without extra cues. Our method generates features that are robust against object occlusion as well as pedestrian interference. Our network also has an exceptional generalized ability that shows good performance even in holistic Re-ID.

5. REFERENCES

[1] Shizhen Zhao, Changxin Gao, Jun Zhang, Hao Cheng, Chuchu Han, Xinyang Jiang, Xiaowei Guo, Wei-Shi Zheng, Nong Sang, and Xing Sun, "Do not disturb me: Person re-identification under the interference of other pedestrians," in *European Conference on Computer Vision*. Springer, 2020, pp. 647–663.

[2] Lingxiao He, Yinggang Wang, Wu Liu, He Zhao, Zhenan Sun, and Jiashi Feng, "Foreground-aware pyramid reconstruction for alignment-free occluded person re-identification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 8450–8459.

[3] Jiaxu Miao, Yu Wu, Ping Liu, Yuhang Ding, and Yi Yang, "Pose-guided feature alignment for occluded person re-identification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 542–551.

[4] Lingxiao He and Wu Liu, "Guided saliency feature learning for person re-identification in crowded scenes," in *European Conference on Computer Vision*. Springer, 2020, pp. 357–373.

[5] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang, "Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 480–496.

[6] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian, "Scalable person re-identification: A benchmark," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1116–1124.

[7] Zhedong Zheng, Liang Zheng, and Yi Yang, "Unlabeled samples generated by gan improve the person re-identification baseline in vitro," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 3754–3762.

[8] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang, "Deepreid: Deep filter pairing neural network for person re-identification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 152–159.

[9] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang, "Bag of tricks and a strong baseline for deep person re-identification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 0–0.

[10] Meng Zheng, Srikrishna Karanam, Ziyan Wu, and Richard J Radke, "Re-identification with consistent attentive siamese networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5735–5744.

[11] Wenjie Yang, Houjing Huang, Zhang Zhang, Xiaotang Chen, Kaiqi Huang, and Shu Zhang, "Towards rich feature discovery with class activation maps augmentation for person re-identification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1389–1398.

[12] Guan'an Wang, Shuo Yang, Huanyu Liu, Zhicheng Wang, Yang Yang, Shuliang Wang, Gang Yu, Erjin Zhou, and Jian Sun, "High-order information matters: Learning relation and topology for occluded person re-identification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 6449–6458.

[13] Kuan Zhu, Haiyun Guo, Zhiwei Liu, Ming Tang, and Jinqiao Wang, "Identity-guided human semantic parsing for person re-identification," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*. Springer, 2020, pp. 346–363.

[14] Mengxi Jia, Xinhua Cheng, Yunpeng Zhai, Shijian Lu, Siwei Ma, Yonghong Tian, and Jian Zhang, "Matching on sets: Conquer occluded person re-identification without alignment," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021, vol. 35, pp. 1673–1681.

[15] Jie Hu, Li Shen, and Gang Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[16] Yulin Li, Jianfeng He, Tianzhu Zhang, Xiang Liu, Yongdong Zhang, and Feng Wu, "Diverse part discovery: Occluded person re-identification with part-aware transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 2898–2907.

[17] Adam Santoro, David Raposo, David GT Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Timothy Lillicrap, "A simple neural network module for relational reasoning," *arXiv preprint arXiv:1706.01427*, 2017.

[18] Hyunjong Park and Bumsub Ham, "Relation network for person re-identification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, vol. 34, pp. 11839–11847.