PROGRESSIVE-GRANULARITY RETRIEVAL VIA HIERARCHICAL FEATURE ALIGNMENT FOR PERSON RE-IDENTIFICATION

*Zhaopeng Dou, Zhongdao Wang, Yali Li, Shengjin Wang**

Department of Electronic Engineering and BNRist, Tsinghua University, China

ABSTRACT

Person re-identification (re-ID) aims to match pedestrian images from non-overlapping cameras. It is a challenging task because of the feature misalignment problem caused by occlusion. In this paper, inspired by the coarse-to-fine nature of human perception, we propose a novel Progressive-Granularity Retrieval (PGR) method to tackle this issue. Specifically, *(i)* we define instance-level, part-level and pixel-level features for an image. PGR learns these features by a single feature extractor to capture hierarchical clues in the image. *(ii)* These features are inherently related but different in perceptual granularity, and they can provide complementary information. For each type of feature, we propose a corresponding similarity metric to achieve hierarchical feature alignment. *(iii)* In training, we learn the model end-to-end. In inference, a progressive retrieval strategy is introduced to efficiently aggregate the complementary information provided by these features. Extensive experiments on three benchmarks of both occluded and holistic-body re-ID tasks show the effectiveness of the proposed method. Especially, our method significantly outperforms state-of-the-art by 4.5% Rank-1 score on the challenging Occluded-Duke dataset.

Index Terms— Person re-identification, progressive-granularity retrieval, feature alignment.

1. INTRODUCTION

Person re-identification (re-ID) aims at retrieving images of a specific person from a large database. It has many practical applications, such as video surveillance, unmanned retail, and smart city. Previous arts [1, 2, 3] show great improvements on datasets where most images depict the holistic human body. However, in real scenes, pedestrians can be easily occluded by various obstacles like cars, bikes and trees [4, 5]. As shown in Fig. 1(a), the uncontrollable occlusions result in the feature misalignment issue, significantly degrading the performance.

Recently, many methods [4, 6, 7, 8] are proposed to learn part features to handle the misalignment issue. Although they have been proved to be effective, there are two limitations: (1) They align two images via part-to-part matching, without considering the relationship between the part and the whole. This makes the retrieval error-prone when the learned part is inaccurate. (2) Part features are still coarse-grained. It may fail to handle the situation where a part itself is partially occluded or has large relative deformation.

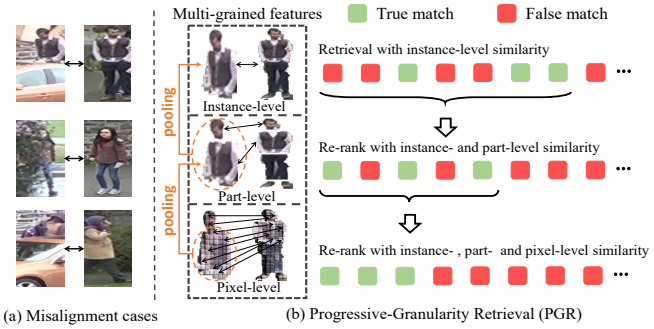

Fig. 1. Motivation and framework of PGR. (a) Cases that contain the feature misalignment issue. (b) Core idea of PGR. For an image, we extract multi-granularity features. For each type of feature, we propose an alignment-based similarity metric (left). During inference, we gradually shrink the gallery list by considering finer-grained feature metric (right).

In this paper, inspired by the "top-down" perception mechanism in the human visual system [9], we propose a Progressive-Granularity Retrieval (PGR) method to tackle these issues via hierarchical feature alignment. Our key idea is illustrated in Fig. 1(b). We use a single model to extract instance-, part- and pixel-level features for an image, which is achieved with the help of a pixel-wise part classifier. These features are inherently related but different in perceptual granularity. Specifically, the part-level feature comes from pooling pixel features inside a part, and the instance-level feature comes from pooling part features in the foreground. For each type of feature, we propose an alignment-based similarity metric, which is calculated by considering the instance/part/pixel-level pedestrian correspondences. These features and metrics can provide complementary information to achieve superior performance. However, finer-grained metric costs more computation, making deployment infeasible. To remedy this issue, we propose a progressive retrieval

This work was supported by the state key development program in 14th Five-Year under Grant No. 2021YFF0602103, 2021YFF0602102, 2021QY1702. We also thank for the research fund under Grant No. 2019GQG0001 from the Institute for Guo Qiang, Tsinghua University.
* is the corresponding author. Emails: wgsgj@tsinghua.edu.cn

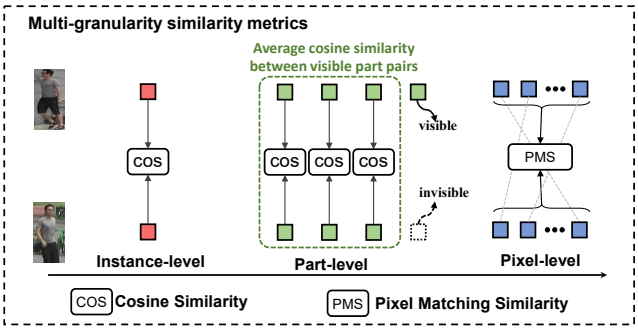

Fig. 2. Similarity metrics. For instance-level metric, we employ the simple cosine similarity. For part-level metric, we only consider similarities between visible parts. For pixel-level metric, we mine the *pixel-pixel* correspondences and average the cosine similarities between the mined pixel-pairs.

strategy: starting from using coarse-grained similarity only, we gradually shrink the gallery list and then add finer-grained similarity to re-rank the remained gallery. This manner efficiently employs the complementary information among these features. Compared with the trivial ensemble, it significantly reduces the computation cost with rarely accuracy drop.

Our contributions are summarized as follows: (1) We define instance-level, part-level and pixel-level features for an image and propose a method to learn these features by a single feature extractor to achieve hierarchical feature alignment. (2) We propose a progressive-granularity retrieval method that efficiently aggregates the complementary information provided by different granularity features. (3) The proposed PGR is a unified method that addresses holistic-body and occluded re-ID simultaneously. Experimental results show that PGR consistently outperforms top-performing methods.

2. PROPOSED METHOD

In this work, we employ a single deep network to extract features in three different granularities and accordingly propose three similarity metrics to achieve hierarchical feature alignment (Section. 2.1). We then discuss the properties of these features and metrics, and finally present a progressive-grained retrieval strategy to efficiently employ complementary information among them (Section 2.2). Fig. 2 shows multi-granularity similarity metrics, and Fig. 1(b) shows the progressive-granularity retrieval strategy.

2.1. Multi-Granularity Features and Metrics

Given an person image $\boldsymbol{I}$, our feature extractor ϕ maps the image into a feature map $\boldsymbol{X} = \phi(\boldsymbol{I}) \in \mathbb{R}^{h \times w \times d}$, where h, w indicates the spatial size and d represents feature dimension.

Instance-level feature and metric. In practical applications, the learned features can be easily influenced by the background and occlusions. For suppressing these negative impacts, we adopt an external human parsing model, SCHP [10], to predict a binary human mask $\boldsymbol{M} \in \{0,1\}^{h \times w}$. Accordingly, we average pixel feature vectors inside the mask as the instance-level feature,

$$\boldsymbol{x}_{\text{ins}} \in \mathbb{R}^d = \frac{1}{\sum_{\forall(i,j)} M(i,j)} \sum\nolimits_{\forall(i,j)} M(i,j)\boldsymbol{X}(i,j) \quad (1)$$

where $\boldsymbol{X}(i,j)$ is the vector in the feature map $\boldsymbol{X}$ at spatial position (i,j). $M(i,j)$ is a binary scalar representing pixel (i,j) belongs to the foreground. The instance-level similarity metric between a query q and a gallery image g is given by,

$$s_{\text{ins}}(q,g) = \frac{1}{2} + \frac{1}{2} \frac{\langle \boldsymbol{x}^q_{\text{ins}}, \boldsymbol{x}^g_{\text{ins}} \rangle}{\|\boldsymbol{x}^q_{\text{ins}}\| \|\boldsymbol{x}^g_{\text{ins}}\|} \quad (2)$$

In the training phase, we apply Softmax Cross-Entropy loss with label smoothing [11] and triplet loss [12] on $\boldsymbol{x}_{\text{ins}}$ for identity-wise classification.

Part-level feature and metric. To learn the part-level feature, we train a pixel-wise part classifier ψ to determine the part to which each pixel on the feature map belongs. Different from previous part-based re-ID methods [13, 14] that employ off-the-shelf human parsing models, in this work, we instead adopt a pseudo-label-based method similar to [7] to learn the classifier, without external part supervision. Specifically, we perform clustering and learning alternately in the training phase. In the clustering stage, we collect foreground pixel feature vectors from all instances of the same identity, and independently perform the K-means [15] algorithm inside the feature sets of each identity, with fixed cluster number p. The pseudo labels of clusters are given by sorting their mean vertical location, labeled as 1 to p from top to bottom. The clustering is performed every 4 epochs. The pseudo part labels are used to train the pixel-wise part classifier ψ, which is shared by all identities. Specifically, ψ takes a feature map $\boldsymbol{X}$ as input and outputs a $(p+1)$-way part probability map $\psi(\boldsymbol{X}) = \boldsymbol{P} \in \mathbb{R}^{h \times w \times (p+1)}$, where the last channel corresponds to the background. The part classifier is trained with Softmax Cross-Entropy loss. We compute each part feature by weighted averaging pixel features in the feature map,

$$\boldsymbol{p}_k \in \mathbb{R}^d = \frac{1}{\sum_{\forall(i,j)} P(i,j,k)} \sum\nolimits_{\forall(i,j)} P(i,j,k)\boldsymbol{X}(i,j) \quad (3)$$

In the training phase, we individually apply Softmax Cross-Entropy loss with label smoothing [11] and triplet loss [12] for each part feature $\boldsymbol{p}_k$.

For the part-level similarity metric, when comparing a query q with a gallery sample g, we only consider parts visible in both samples. The visibility of a part is reasoned by checking if there exists at least a single pixel categorized into this part by the part classifier ψ. Let $v^q_k, v^g_k \in \{0,1\}$ denote the binary visibility of the k-th part of q and g respectively. The part-level similarity between q and g is given by,

$$s_{\text{part}}(q,g) = \frac{\sum_{k=1}^{p} (v^q_k \cdot v^g_k)\, s_{\text{part}}(q,g,k)}{1 + \sum_{k=1}^{p} v^q_k \cdot v^g_k} \quad (4)$$

where $s_{\text{part}}(q,g,k) = \frac{1}{2} + \frac{1}{2}\frac{\langle \boldsymbol{p}_k^q, \boldsymbol{p}_k^g \rangle}{\|\boldsymbol{p}_k^q\| \|\boldsymbol{p}_k^g\|}$ is the cosine distance between the k-th part features.

Note: We can obtain the foreground mask by $\boldsymbol{M} = \sum_{k=1}^{p} \boldsymbol{P}(:,:,k)$. This allows us to discard the human parsing model during inference, making inference more efficient.

Pixel-level feature and metric. Part-based features may dramatically change when the corresponding parts are occluded or deformed. To capture even finer visual invariance across different instances of the same identity, we consider pixel-level similarity. Specifically, we first flatten the foreground pixel features of the query q and the gallery sample g into sets $\mathcal{X}^q = \{\boldsymbol{x}_1^q, \boldsymbol{x}_2^q, \ldots, \boldsymbol{x}_m^q\}$ and $\mathcal{X}^g = \{\boldsymbol{x}_1^g, \boldsymbol{x}_2^g, \ldots, \boldsymbol{x}_n^g\}$, respectively. Then we mine the *pixel-pixel* correspondences between sets $\mathcal{X}^q$ and $\mathcal{X}^g$ under the cycle consistency constraint. Let s_{ij} denotes the cosine similarity between the $\boldsymbol{x}_i^q$ and $\boldsymbol{x}_j^g$. We can find the index of the maximum response of $\boldsymbol{x}_i^q$ in $\mathcal{X}^g$ by $j^* = \arg\max_j s_{ij}$. Then we find the maximum response of $\boldsymbol{x}_{j*}^g$ in $\mathcal{X}^q$ by $i^* = \arg\max_i s_{ij*}$. The spatial distance between the positions of $\boldsymbol{x}_i^q$ and $\boldsymbol{x}_{i*}^q$ in the original feature map is denoted by d_i, then the mined pixel correspondences set is,

$$\mathcal{P} = \{(\boldsymbol{x}_i^q, \boldsymbol{x}_{j*}^g) | d_i < \epsilon\} \tag{5}$$

where ϵ is a threshold controlling the hardness of mining pixel correspondences. The pixel-level similarity between q and g is formulated by:

$$s_{\text{pixel}}(q,g) = \frac{1}{2} + \frac{1}{2|\mathcal{P}|} \sum\nolimits_{(\boldsymbol{x}_i^q, \boldsymbol{x}_{j*}^g) \in \mathcal{P}} \frac{\langle \boldsymbol{x}_i^q, \boldsymbol{x}_{j*}^g \rangle}{\|\boldsymbol{x}_i^q\| \|\boldsymbol{x}_{j*}^g\|} \tag{6}$$

where $|\mathcal{P}|$ is the number of elements in $\mathcal{P}$.

2.2. Progressive-Granularity Retrieval

Multi-granularity features are complementary. We show the complementarity with a qualitative retrieval example in Fig. 3 (a-d). First, the instance-level similarity roughly reflects how two samples are alike "at a glance". It may fail in distinguishing local differences. Second, the part-level metric considers visible part pairs only, so it is less affected by occlusion. However, it may fail when a part itself is partially visible or has a large deformation. At last, the pixel-level metric can capture finer similarities, even between partially visible parts. The first gallery image in Fig. 3 (c) is an example, where the upper body is a part and half of it is occluded. This sample is missed in retrieval with instance-level and part-level features but successfully recalled with pixel-level features.

We further validate the complementarity with quantitative results. First, we evaluate the Mean Ranking Difference (MRD) between retrieval results with two individual features. MRD is defined as the mean absolute difference of rankings of all positive samples in the gallery, *e.g.*, a gallery sample ranks at 5^{th} with the instance-level feature while ranks at 7^{th} with the part-level feature, then the difference value is 2. We perform experiments on a toy test set Occluded-REID [5] and

(e) Mean Ranking Difference (MRD)

Instance-part	Part-pixel	Instance-pixel
20.83	23.56	29.64

(f) Performance and time cost

No.	Inst.	Part	Pixel	Prog.	R@1	Time cost
I	✓				60.6	43 ms
II		✓			68.1	217 ms
III			✓		75.4	4532 s
IV	✓	✓	✓		80.6	4578 s
V	✓	✓	✓	✓	80.4	130 s

Fig. 3. Analysis of the complementarity and efficiency. (a-e) Retrieval results. (The green box indicates a true match) (f) Performance and time cost on Occluded-REID [5].

show results in Fig. 3 (e). The MRD between individual features is significantly large (> 20), showing their differences in rankings. Second, we test the retrieval accuracy of each feature respectively, and also the accuracy of aggregating multi-grained features. Results are shown in Fig. 3 (f). We observe that each feature present decent accuracy (I-III), while the ensemble shows further improvements (IV). Combined, multi-granularity features are indeed complementary.

Progressive-granularity retrieval. Intuitively, finer-granularity similarities require more complicated computation. Fig. 3(f) shows the time cost using different similarity metrics in retrieval. We observe that the computation of instance-level and part-level similarities are light while the cost of pixel-level similarity is significantly more expensive. To efficiently utilize the complementary information from multi-granularity features, we propose a progressive retrieval manner. We first use the instance-level similarity s_{ins} to rank the gallery. Then we select the top-N_1 results to form a new gallery and use $s_1 = \alpha_1 s_{\text{ins}} + (1-\alpha_1) s_{\text{part}}$ to re-rank it, where α_1 is a hyper-parameter. We further select the top-N_2 ($N_2 < N_1$) results as a new gallery and use $s_2 = \alpha_2 s_1 + (1-\alpha_2) s_{\text{pixel}}$ to re-rank it to obtain the final results. Since the major computation bottleneck is the pixel-level similarity, the speedup ratio *w.r.t.* trivial ensemble is mainly related to the ratio between the original gallery size and final gallery size, *i.e.*, $\frac{|\mathcal{G}|}{N_2}$. Compared with trivial ensemble, PGR significantly speeds up the retrieval process by 35× with rarely accuracy drop (see IV and V in Fig. 3(f)).

3. EXPERIMENTS

3.1. Datasets and Implementation Details

Datasets and protocols. Experiments are conducted on Market-1501 [16], DukeMTMC-reID [17] and Occluded-Duke [4]. Occluded-Duke is more challenging due to the severe misalignment issue caused by occlusion. Following [6], the Cumulative Matching Characteristic (CMC) and mean average precision (mAP) are used as evaluation metrics. All experiments are performed in single query setting.

Table 1. Performance comparison on Occluded-Duke [4]

Methods	Backbone	Occluded-Duke	
		R@1	mAP
PCB+RPP [8]	ResNet50	42.6	33.7
SFR [21]	ResNet50	42.3	32.0
PGFA [4]	ResNet50	51.4	37.3
HOReID [6]	ResNet50	55.1	43.8
PGR (Ours)	ResNet50	**62.8**	**50.1**
ISP [7]	HRNet-W32	62.8	52.3
PAT [22]	transformer	64.5	53.6
PGR (Ours)	HRNet-W32	**69.0**	**57.4**

Table 2. Performance comparison on holistic-body datasets: Market-1501 [16] and DukeMTMC-reID [17].

Methods	Backbone	Market-1501		Duke	
		R@1	mAP	R@1	mAP
PCB+RPP [8]	ResNet50	92.3	77.4	81.8	66.1
PGFA [4]	ResNet50	91.2	76.8	82.6	65.5
SSP-ReID [23]	ResNet50	92.5	75.8	81.8	68.9
HOReID [6]	ResNet50	94.2	84.9	86.9	75.6
PGR (Ours)	ResNet50	**94.8**	**85.3**	**87.1**	**76.5**
ISP [7]	HRNet-W32	95.3	88.6	89.6	80.0
PAT [22]	transformer	95.4	88.2	88.8	78.2
PGR (Ours)	HRNet-W32	**95.8**	**89.3**	**90.9**	**81.0**

Implementation Details. For fair comparison, we adopt ResNet50 [18] and HRNet-W32 [19] as the backbone, respectively. Each mini-batch contains 32 images from 8 different identities. The input image is resized to 256×128 and augmented by random cropping, horizontal flopping and random erasing [20]. The ϵ in Eq. 5 and α_1, α_2, N_1 and N_2 in Sec. 2.2 are empirically set as 4, 0.05, 0.8, 1000 and 30, respectively. The number of semantic parts in Eq. 4 is set as 5.

3.2. Comparisons with the State-of-the-art Methods

We compare PGR with prior arts on three benchmarks, *i.e.*, Occluded-Duke, Market-1501 and DukeMTMC-reID. Results are shown in Table 1 and Table 2. We can make several observations: (1) PGR consistently shows superior performance on these three benchmarks. This demonstrates that PGR is capable on both holistic-body re-ID and occluded re-ID tasks. (2) When employing ResNet50 [18] as the backbone, on Occluded-Duke, PGR outperforms previous state-of-the-art method HOReID [6] by 7.7% Rank-1 score. When adopting HRNet-W32 [7] as the backbone, PGR exceeds ISP [7] using the same backbone by 6.2% Rank-1 score. This shows that PGR is robust for different backbones. (3) Although PAT [22] uses the advanced architecture, *i.e.*, transformer [24], PGR shows better performance than it, further showing the effectiveness of the proposed PGR.

3.3. Ablation study

We verify the effectiveness of each type of feature here, and results are shown in Table. 3. The baseline is the IDE model [25], which pooling the entire feature map as the final representation, without considering the feature misalignment issue. In our method, when using the instance-level feature only, the Rank-1 score is improved by 5.2%. when additionally using the part-level and pixel-level features, the Rank-1 score is further improved by 5.7% and 4.5%, respectively. We can draw two conclusions. First, the multi-granularity features and metrics are all effective and they can achieve feature alignment at different perceptual granularities. Second, these features and metrics can provide complementary information, and thus enable superior performance after aggregation.

Table 3. Effectiveness of each type of feature on Occluded-Duke with ResNet50 backbone.

Methods	Feature			Occluded-Duke	
	Inst.	Part.	Pixel	R@1	mAP
Baseline	–	–	–	47.4	41.6
PGR-I	✓	–	–	52.6	45.3
PGR-II	✓	✓	–	58.3	48.6
PGR-III	✓	✓	✓	**62.8**	**50.1**

3.4. Visualization

Fig. 4(a) shows the learned parts, in which different colors represent different semantic parts and the shaded area represents the background. We can see that for an identity, the background noise is effectively eliminated and the specific semantic parts in different images are consistent. This shows that we can achieve feature alignment at instance-level and part-level. Fig. 4 (b) shows the mined pixel correspondences. We can see that PGR ignores the irrelevant pixels and aligns two images at pixel-level. The right two images in Fig. 4(b) are an example: when calculating the similarity, we focus on pixel pairs of the upper body appearing in two images, and ignore the lower body that only appears in the right image.

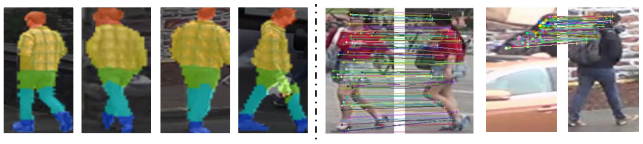

Fig. 4. Visualization of the hierarchical feature alignment. (a) the learned parts. (b) the mined pixel correspondences.

4. CONCLUSION

Inspired by the coarse-to-fine nature of human perception, we propose a Progressive-Granularity Retrieval (PGR) method for both holistic-body and occluded re-ID tasks. We extract instance-level, part-level and pixel-level features of persons with a single appearance model. For each type of feature, we propose a corresponding similarity metric to achieve hierarchical feature alignment. The complementary information among these features is efficiently aggregated by a progressive strategy. Finally, extensive experiments demonstrate the effectiveness of the proposed method.

5. REFERENCES

[1] Xuesong Chen, Canmiao Fu, Yong Zhao, Feng Zheng, Jingkuan Song, Rongrong Ji, and Yi Yang, "Salience-guided cascaded suppression network for person re-identification," in *CVPR*, 2020, pp. 3300–3310.

[2] Dongkai Wang and Shiliang Zhang, "Unsupervised person re-identification via multi-label classification," in *CVPR*, 2020, pp. 10981–10990.

[3] Yukun Huang, Zheng-Jun Zha, Xueyang Fu, Richang Hong, and Liang Li, "Real-world person re-identification via degradation invariance learning," in *CVPR*, 2020, pp. 14084–14094.

[4] Jiaxu Miao, Yu Wu, Ping Liu, Yuhang Ding, and Yi Yang, "Pose-guided feature alignment for occluded person re-identification," in *ICCV*, 2019, pp. 542–551.

[5] Jiaxuan Zhuo, Zeyu Chen, Jianhuang Lai, and Guangcong Wang, "Occluded person re-identification," in *ICME*. IEEE, 2018, pp. 1–6.

[6] 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 *CVPR*, 2020, pp. 6449–6458.

[7] Kuan Zhu, Haiyun Guo, Zhiwei Liu, Ming Tang, and Jinqiao Wang, "Identity-guided human semantic parsing for person re-identification," in *ECCV*. Springer, 2020, pp. 346–363.

[8] 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 *ECCV*, 2018, pp. 480–496.

[9] Timothy J Buschman and Earl K Miller, "Top-down versus bottom-up control of attention in the prefrontal and posterior parietal cortices," *science*, vol. 315, no. 5820, pp. 1860–1862, 2007.

[10] Peike Li, Yunqiu Xu, Yunchao Wei, and Yi Yang, "Self-correction for human parsing," *IEEE TPAMI*, 2020.

[11] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna, "Rethinking the inception architecture for computer vision," in *CVPR*, 2016, pp. 2818–2826.

[12] Alexander Hermans, Lucas Beyer, and Bastian Leibe, "In defense of the triplet loss for person re-identification," *arXiv preprint arXiv:1703.07737*, 2017.

[13] Mahdi M Kalayeh, Emrah Basaran, Muhittin Gökmen, Mustafa E Kamasak, and Mubarak Shah, "Human semantic parsing for person re-identification," in *CVPR*, 2018, pp. 1062–1071.

[14] Houjing Huang, Xiaotang Chen, and Kaiqi Huang, "Human parsing based alignment with multi-task learning for occluded person re-identification," in *ICME*. IEEE, 2020, pp. 1–6.

[15] James MacQueen et al., "Some methods for classification and analysis of multivariate observations," in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*. Oakland, CA, USA, 1967, vol. 1, pp. 281–297.

[16] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian, "Scalable person re-identification: A benchmark," in *ICCV*, 2015, pp. 1116–1124.

[17] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in *ECCV*. Springer, 2016, pp. 17–35.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[19] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang, "Deep high-resolution representation learning for human pose estimation," in *CVPR*, 2019, pp. 5693–5703.

[20] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang, "Random erasing data augmentation.," in *AAAI*, 2020, pp. 13001–13008.

[21] Lingxiao He, Zhenan Sun, Yuhao Zhu, and Yunbo Wang, "Recognizing partial biometric patterns," *arXiv preprint arXiv:1810.07399*, 2018.

[22] 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 *CVPR*, June 2021, pp. 2898–2907.

[23] Rodolfo Quispe and Helio Pedrini, "Improved person re-identification based on saliency and semantic parsing with deep neural network models," *Image and Vision Computing*, vol. 92, pp. 103809, 2019.

[24] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *NeurIPS*, 2017, pp. 5998–6008.

[25] Liang Zheng, Yi Yang, and Alexander G Hauptmann, "Person re-identification: Past, present and future," *arXiv preprint arXiv:1610.02984*, 2016.