# Semantically Enhanced Attention Map-Driven Occluded Person Re-identification

Yiyuan Ge[1], Mingxin Yu[1], Zhihao Chen[1], Wenshuai Lu[2], and Huiyu Shi[2]

[1]Beijing Information Science and Technology University
[2]Tsinghua University

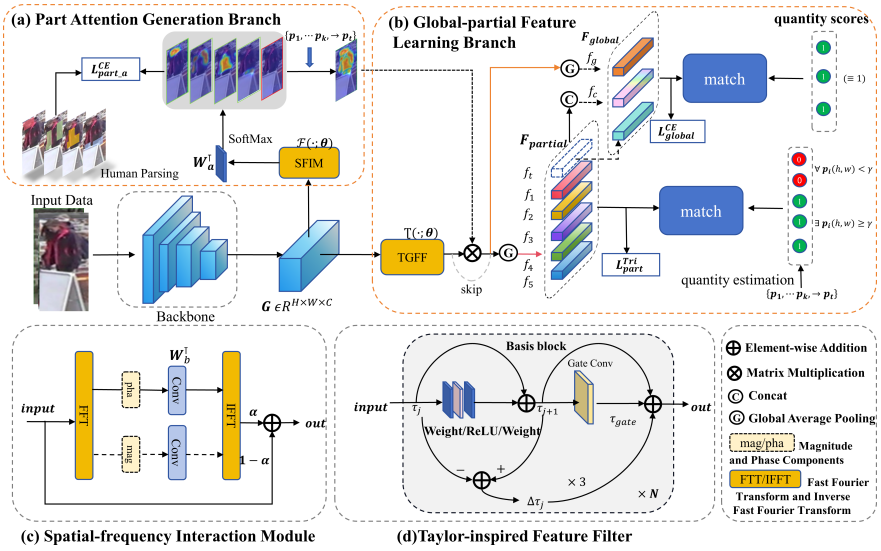
May 02, 2024

## Abstract

Occluded person Re-identification (Re-ID) is to identify a particular person when the person's body parts are occluded. However, challenges remain in enhancing effective information representation and suppressing background clutter when considering occlusion scenes. In this paper, we propose a novel Attention Map-Driven Network (AMD-Net) for occluded person Re-ID. In AMD-Net, human parsing labels are introduced to supervise the generation of partial attention maps, while we suggest a Spatial-frequency Interaction Module (SIM) to complement the higher-order semantic information from the frequency domain. Furthermore, we propose a Taylor-inspired Feature Filter (TFF) for mitigating background disturbance and extracting fine-grained features. Moreover, we also design a part-soft triplet loss, which is robust to non-discriminative body partial features. Experimental results on Occluded-Duke, Occluded-Reid, Market-1501, and Duke-MTMC datasets show that our method outperforms existing state-of-the-art methods. The code is available at: https://github.com/ISCLab-Bistu/SA-ReID.

## Hosted file

`Main Document.docx` available at https://authorea.com/users/777128/articles/897954-semantically-enhanced-attention-map-driven-occluded-person-re-identification



# Semantically Enhanced Attention Map-Driven Occluded Person Re-identification

Yiyuan Ge[1], Mingxin Yu[1], Zhihao Chen[2], Wenshuai Lu[3], and Huiyu Shi[3]

[1] *School of Instrument Science and Opto-Electronics Engineering, Beijing Information Science and Technology University, Beijing, China*

[2] *School of Computer, Beijing Information Science and Technology University, Beijing, China*

[3] *Department of Precision Instrument, Tsinghua University, Beijing, China*

Email: yumingxin@bistu.edu.cn.

Occluded person Re-identification (Re-ID) is to identify a particular person when the person's body parts are occluded. However, challenges remain in enhancing effective information representation and suppressing background clutter when considering occlusion scenes. In this paper, we propose a novel Attention Map-Driven Network (AMD-Net) for occluded person Re-ID. In AMD-Net, human parsing labels are introduced to supervise the generation of partial attention maps, while we suggest a Spatial-frequency Interaction Module (SIM) to complement the higher-order semantic information from the frequency domain. Furthermore, we propose a Taylor-inspired Feature Filter (TFF) for mitigating background disturbance and extracting fine-grained features. Moreover, we also design a part-soft triplet loss, which is robust to non-discriminative body partial features. Experimental results on Occluded-Duke, Occluded-Reid, Market-1501, and Duke-MTMC datasets show that our method outperforms existing state-of-the-art methods. The code is available at: https://github.com/ISCLab-Bistu/SA-ReID.

*Introduction:* The purpose of person Re-identification (Re-ID) is to identify different instances across cameras and viewpoints [1-3]. With the advances in deep learning and intelligent algorithms [4], Re-ID technology has been successfully applied to smart security, person search, and other fields. However, occlusion occurs from time to time in real person Re-ID scenarios, which severely impairs model performance. Under occlusion conditions, the original images contain fewer valid features as well as more occlusion noise, which leads to the failure of instances matching.

In order to address the above issues, some methods [5], [6] use a pose estimator to extract features around keypoints, but this method is not stable for the person Re-ID datasets due to domain differences. Recently, several part-based approaches have been proposed and demonstrated excellent performance [7], [8]. These methods attempt to adaptively construct body partial feature representations and combine the global feature representations to suppress the effects of occlusion.

However, we believe that the current approach still suffers from the following limitations. **Firstly**, due to the lack of human topological prior, the part-based approaches could not accurately constrain the regions of local feature pooling, which leads to low-quality partial feature representations. **Secondly**, in occlusion scenes, the effective information is limited. There is no specific solution on how to reasonably increase the effective information. In addition, background clutter remains a serious distraction due to the lack of a special fine-grained feature filter. **Thirdly**, the prevailing part-based methods [9], [10] currently rely on identity loss and hard triplet loss to supervise individual partial feature blocks for learning discriminative representations. However, this approach lacks reasonability, as two people with different identities may have remarkably similar appearances in specific body parts.

In response to the above problems, this letter proposes specific solutions. Our contributions are summarised below:

(1) We propose a novel approach for occluded person Re-ID named Attention Map-Driven Network (AMD-Net). We use human pasing labels and a Spatial-frequency Interaction Module (SIM) to generate semantically enhanced body partial attention maps.

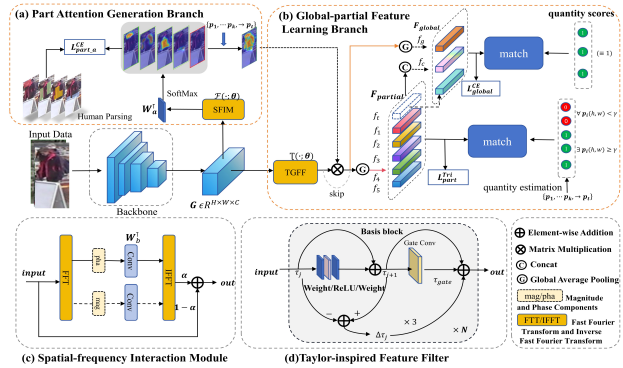



**Fig. 1** *The pipeline of ACM-Net. It consists of the part attention generation branch and the global-partial feature learning branch.*

(2) Taylor-inspired Feature Filter (TFF) is designed to suppress background interference and extract effective features. Combine with semantically enhanced attention maps, AMD-Net can generate global and partial features with finer granularity.

(3) We introduce a part-soft triplet loss to supervise body partial features, which is robust to occlusion and similar body part appearance.

*Part Attention Generation Branch:* Part Attention Generation Branch (PAGB) is shown in Figure 1(a). Body partial attention maps are subject to dual supervision, one is the human parsing labels, which are generated by a pre-trained pose estimation network [11], and the other is Re-ID loss, which includes identity loss and triplet loss. Human parsing labels provide a restricted region for each body partial attention, and then Re-ID loss supervises the attention maps to focus on identity-related feature parts. In addition, we design the SIM to aggregate high-level semantic information in the frequency domain.

Let $\boldsymbol{G} \in \mathrm{R}^{H\times W\times C}$ represent the input tensor of the PAGB, which is extracted by the backbone network. Firstly, SIM is used for domain information enhancement. Then, we employ a $1\times 1$ convolutional layer followed by the SoftMax function to derive the attention maps $\boldsymbol{P} \in \mathrm{R}^{H\times W\times k}$: $\boldsymbol{P} = \{\boldsymbol{p_1}, \cdots \boldsymbol{p_k}\}$. $\{\boldsymbol{p_1}, \cdots \boldsymbol{p_k}\}$ represent the attention maps of $k$ body parts, respectively. In this paper, the value of $k$ is 5, which denotes the body regions of the head, arms, trunk, legs, and feet. The entire process could be expressed as follows:

$$\boldsymbol{P} = \mathrm{softmax}(\boldsymbol{\mathcal{F}}(\boldsymbol{G};\boldsymbol{\theta}) \cdot \boldsymbol{W_a}^{\boldsymbol{T}}) \tag{1}$$

where $\boldsymbol{W_a}$ is the weight of the $1\times 1$ convolution. $\boldsymbol{\mathcal{F}}(\cdot;\boldsymbol{\theta})$ represents the process of frequency-space domain information interaction, and $\boldsymbol{\theta}$ is the process parameter. Pixel value $\boldsymbol{p_i}(h,w)$ of $i$-th map represents the probability that it belongs to the $i$-th body part. We calculate the loss of partial attention:

$$L_{part\_a}^{CE} = -\sum_{k=1}^{K}\sum_{h=0}^{H-1}\sum_{w=0}^{W-1} q_k \cdot log(\boldsymbol{p_k}(h,w)),$$
$$q_k = \begin{cases} 1-\sigma+\dfrac{\sigma}{N} & if\ \boldsymbol{H}(h,w)=k \\ \dfrac{\sigma}{N} & otherwise \end{cases} \tag{2}$$

where $N$ is the batchsize, $\boldsymbol{H} \in \mathrm{R}^{H\times W\times 1}$ represents the human parsing label. If the pixel position $(h,w)$ belongs to the $k$ body parts, the pixel value $\boldsymbol{H}(h,w)$ is assigned as $\{1,\cdots k\}$; otherwise, it is set to 0 for the background. $\sigma$ is the label smoothing coefficient, which is set to 0.15 in this paper.

*Spatial-frequency Interaction Module:* Previous studies [12] have shown that the phase component of the Fourier transform usually preserves higher-order semantics of the original signal, while the amplitude component includes low-level modal information. Most existing Re-ID methods [5-8] rely on spatial domain processing to extract effective features while omitting the global information within the Fourier domain. SIM is proposed to aggregate the high-level semantic information contained in the Fourier domain. As depicted in

Figure 1(c), we apply the fast Fourier transform to acquire the phase and amplitude components. Then, we introduce the convolutional induction bias in the Fourier domain to enhance the generalisation ability of the model. Finally, we interact information in the spatial and frequency domains to generate semantically enhanced attention maps. The key operations of SIM can be summarized as follows:

$$\boldsymbol{F^{mag}},\ \boldsymbol{F^{pha}} = \mathcal{F}_{\mathrm{FFT}}(\boldsymbol{G};\boldsymbol{\theta}) \tag{3}$$

$$\boldsymbol{F^{out}} = \alpha \cdot [\mathcal{F}_{\mathrm{IFFT}}(\boldsymbol{F^{pha}} \cdot \boldsymbol{W_b}^{\boldsymbol{T}};\boldsymbol{\theta})] + (1-\alpha)\cdot \boldsymbol{G} \tag{4}$$

where $\mathcal{F}_{\mathrm{FFT}}(\cdot;\boldsymbol{\theta})$ and $\mathcal{F}_{\mathrm{IFFT}}(\cdot;\boldsymbol{\theta})$ represent the fast Fourier transform and its corresponding inverse process. $\boldsymbol{F^{mag}}$ and $\boldsymbol{F^{pha}}$ are the amplitude and phase components. $\boldsymbol{W_b}^{\boldsymbol{T}}$ is the weight of the convolution operation. $\alpha$ is the hyper-parameter regulating the interaction between the spatial and frequency domains, which is set to 0.35 in this paper.

*Global-partial Feature Learning Branch:* Within the Global-partial Feature Learning Branch (GFLB), as depicted in Figure 1(b), we first use TFF to extract fine-grained information and suppress background interference, which will be described in next section. The feature after TFF is denoted as $\boldsymbol{F}\epsilon \mathrm{R}^{H\times W\times C}$, and we use the partial attention maps $\boldsymbol{P} = \{\boldsymbol{p_1},\cdots \boldsymbol{p_k}\}$ to generate the foreground attention $\boldsymbol{p_t} \in R^{H\times W\times 1}$ : $\boldsymbol{p_t}(\boldsymbol{h},\boldsymbol{w}) = \mathrm{Max}\{\boldsymbol{p_1}(h,w),\cdots \boldsymbol{p_k}(h,w)\}$. Then, we generate foreground and the $k$ body partial representations as follows:

$$\boldsymbol{F} = \mathcal{T}(\boldsymbol{G};\boldsymbol{\theta}) \tag{5}$$

$$\boldsymbol{f_i} = \frac{\sum_{h=0}^{H-1}\sum_{w=0}^{W-1}\boldsymbol{F}(h,w)\cdot \boldsymbol{p_i}(h,w)}{H\cdot W},\ \ i\epsilon\{t,1,2,\cdots k\} \tag{6}$$

where $\mathcal{T}(\cdot;\boldsymbol{\theta})$ represents the processing of TGFF. Furthermore, we perform a global average pooling operation on $\boldsymbol{F}$ to obtain $\boldsymbol{f_g}$ and concat all body partial features to obtain $\boldsymbol{f_c}$. The process is shown below:

$$\boldsymbol{f_g} = \frac{\sum_{h=0}^{H-1}\sum_{w=0}^{W-1}\boldsymbol{F}(h,w)}{H\cdot W},\ \ \boldsymbol{f_c} = \boldsymbol{f_1}\oplus \boldsymbol{f_2}\cdots \oplus \boldsymbol{f_k} \tag{7}$$

Finally, we use $\boldsymbol{F_{global}} = \{\boldsymbol{f_t},\boldsymbol{f_c},\boldsymbol{f_g}\}$ to denote the global feature representations and $\boldsymbol{F_{part}} = \{\boldsymbol{f_1},\cdots \boldsymbol{f_k}\}$ for the partial feature representations. To mitigate the impact of occluded parts on the model, we employ partial attention maps to generate quality scores that reflect the visibility of body parts. When there exists at least one pixel value in $\boldsymbol{p_i}(i\in\{1,2,\cdots k\})$ that is higher than the threshold $\gamma$ (which is set to 0.4), the quality score of the $i$-th body part is set to 1. In the inference phase, we compare only the visible parts.

*Taylor-inspired Feature Filter:* In the occlusion scenes, person Re-identification task frequently encounters challenges such as background clutter and occlusion noise, which make it difficult for the model to capture fine-grained and effective features. To tackle this problem, we rethink the similarity between ordinary differential equations (ODEs) and residual networks. In prior works [13], the forward Euler was mapped to the residual block and utilized the Eulerian approach to construct a more detailed feature extraction module. Since the Taylor finite difference method has higher accuracy than the Eulerian method in solving numerical ODEs [14], inspired by the Taylor method, we construct a specialised feature filter TFF using residual blocks and gated convolutions (shown in Figure 1(d)). Concretely, we discretize the ODEs using second-order Taylor finite difference equations, and the partial derivatives can be approximated as follows:

$$\frac{\partial_\tau}{\partial_x} = \frac{-3\tau(x)+4\tau(x+h)-\tau(x+2h)}{2h} \tag{8}$$

The above equation can be expressed as:

$$f(\tau_j,x_j) = \frac{-\frac{1}{2}\tau_{j+2}+2\tau_{j+1}-\frac{3}{2}\tau_j}{h} \tag{9}$$

Equivalent variations of the Eq. (9) is as follows:

$$\tau_{j+2} = \tau_{j+1} + [3(\tau_{j+1}-\tau_j)] - 2f(\tau_j,x_j)h \tag{10}$$

Similar to [13], we use multiple convolutional layers to implement the mapping from $\tau_j$ to $\tau_{j+1}$, and $(\tau_{j+1}-\tau_j)$ denotes the residual feature $\Delta\tau_j$. In addition, to suppress the background clutter, we apply the gated convolution $\tau_{gate}$ to obtain $-2f(\tau_j,x_j)h$. This process can be formulated as follows:

$$\tau_{j+2} = \tau_{j+1} + \tau_{gate} + 3\Delta\tau_j \tag{11}$$

Above is the construction process of the basis block in TFF. By superimposing the basis block, TFF enables the extraction of more detailed and intricate feature information from multiple layers. Besides, through the soft gating mechanism in gated convolution, TFF can effectively suppress background interference. We use three basis blocks to build TFF in our AMD-Net.

*Part-Soft Triplet Loss:* Previous occluded person Re-ID methods mainly apply identity loss and hard triplet loss [15] to each individual body partial feature. However, it is important to note that different individuals may exhibit high similarity in certain body parts that lack sufficient distinctiveness for identity discrimination.

In this paper, we design a part-soft triplet loss to enhance the model's capability in handling non-discriminative partial features and improve its overall robustness. we combine the average distance of all body partial features to compute the triplet loss. Particularly, consider that human head contains more discriminative features, such as face [16]. We provide a complementary coefficient $\beta_{head}$ to emphasise the distinction of head feature, which is set to 0.5 in our method. The process of calculating the average distance can be described as follows:

$$d_{(a,b)} = \frac{(1+\beta_{head})\cdot \mathrm{d_{eu}}(f_1^a,f_1^b)+\sum_{k=2}^{K}\mathrm{d_{eu}}(f_k^a,f_k^b)}{K+\beta_{head}} \tag{12}$$

$f_k^{(a)}$ and $f_k^{(b)}$ are the sampled features of two different instances on the k-th body part. $\mathrm{d_{eu}}(\cdot)$ denotes the computation process of the Euclidean distance. For the anchored sample $a$, we compute the hardest positive distance $d_{(a,p)}$ and the hardest negative distance $d_{(a,n)}$, which is similar to [7], [8]. Finally, part-soft triplet loss can be formulated as:

$$L_{part}^{tri} = \sum_{(a,p),(a,n)}\left[d_{(a,p)} - d_{(a,n)} + M\right]_+ \tag{13}$$

where $[\cdot]_+$ stands for hinge loss and $M$ is the distance margin. As mentioned above, our proposed part-soft triplet loss globally optimizes the distances between the corresponding partial features in a softer way. And it enables the model to prioritize discriminative partial features during the training process, while being robust to non-discriminative partial features as well as occlusion features. For global feature $\{\boldsymbol{f_t},\boldsymbol{f_c},\boldsymbol{f_g}\}$, we compute the cross-entropy loss $L_{global}^{CE}$ with label smoothing, which is similar to Eq. (2). Finally, the whole optimisation objective can be described as:

$$L_{all} = L_{part}^{tri} + L_{global}^{CE} + L_{part\_a}^{CE} \tag{14}$$

*Datasets and Implementation Details:* We conduct experiments on the dedicated occluded person Re-ID datasets Occluded-Duke [17] and Occluded-Reid [18] as well as the regular person Re-ID datasets Market-1501 [19] and DukeMTMC [20]. Similar to [1], we adopt mean Average Precision (mAP) and Rank-1 as evaluation metrics.

AMD-Net is built based on the TorchReID framework [21], and is trained and evaluated on 2 RTX 3090 GPUs. We used HRNet-W32 after pre-trained on ImageNet dataset as the backbone network. All images are reshaped to 384×128 and the batchsize is set to 64. We train our model with Adam optimizer for 120 epochs, the initial learning rate is $3.0\times 10^{-4}$, and at the 50th and 80th epoch, it drops to the $3.0\times 10^{-5}$ and $3.0\times 10^{-6}$, respectively.

*Experimental results:* We compare our proposed AMD-Net with existing methods on occluded and regular person Re-ID datasets. As shown in Table 1, AMD-Net outperforms the state-of-the-art methods. In addition,

Table 2 reports the ablation experiments on Occluded-Duke dataset, and the experimental results demonstrate the effectiveness of our design components and loss function.

***Table 1.*** *Compare with state-of-the-art methods on Occluded-Duke, Occluded-Reid, Market-1501, and Duke-MTMC datasets*

| Method | Occluded datasets | | | | Regular datasets | | | |
|---|---|---|---|---|---|---|---|---|
| | Occ-Duke | | Occ-reID | | Market-1501 | | DukeMTMC | |
| | Rank1 | mAP | Rank1 | mAP | Rank1 | mAP | Rank1 | mAP |
| PAT [22] | 64.5 | 53.6 | 81.6 | 72.1 | 95.4 | 88.0 | 88.8 | 78.2 |
| SGR [1] | 69.0 | 57.2 | 78.5 | 72.9 | 96.1 | 89.3 | 91.1 | 81.3 |
| FED [9] | 68.1 | 56.4 | 86.3 | 79.3 | 95.0 | 86.3 | 89.4 | 78.0 |
| FRT [10] | 70.7 | 61.3 | 80.4 | 71.0 | 95.5 | 88.1 | 90.5 | 81.7 |
| QPM [8] | 66.7 | 53.3 | - | - | - | - | - | - |
| HCGA[7] | 70.2 | 57.5 | 87.2 | **95.6** | 95.2 | 88.4 | - | - |
| CAAO [2] | 68.5 | 59.5 | 87.1 | 83.4 | 95.3 | 88.0 | 89.8 | 80.9 |
| **Ours*** | **75.8** | **63.3** | **87.3** | 87.4 | **96.2** | **89.4** | **91.3** | **82.7** |

***Table 2.*** *Ablation experiments on the Occluded-Duke dataset. w/o denotes as "without", r/w denotes as "replace with"*

| Method | | Rank1 | mAP |
|---|---|---|---|
| w/o Part Attention | / | 63.2 | 51.2 |
| w/o SIM | / | 72.3 | 61.2 |
| w/o TFF | / | 73.1 | 60.8 |
| w/o SIM & TFF | / | 70.7 | 58.9 |
| w/o part-soft triplet Loss | r/w ID Loss | 71.5 | 59.8 |
| w/o part-soft triplet Loss | r/w hard-triplet Loss | 72.6 | 60.5 |
| **Ours*** | / | **75.8** | **63.3** |

*Conclusion:* In this paper, we propose an Attention Map-Driven Network (AMD-Net) for occluded person Re-ID. We have improved the shortcomings of the previous approaches in three ways. To begin with, human parsing labels are utilized to establish more precise feature extraction regions. Subsequently, we introduce the Spatial-frequency Interaction Module (SIM) and the Taylor-inspired Feature Filter (TFF) to add valid information and suppress background clutter. Lastly, we suggest a part-soft triplet loss to increase the model's inclusiveness of the non-discriminative body partial features.The comprehensive experimental results on four datasets provide strong evidence of the exceptional performance of AMD-Net.

*Acknowledgments:*This work was supported by the National Natural Science Foundation of China (NSFC) (grant no. U21A6003).



**References**

1. C. Yan, G. Pang, J. Jiao, X. Bai, X. Feng and C. Shen, "Occluded Person Re-Identification with Single-scale Global Representations," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, QC, Canada, 2021, pp. 11855-11864.
2. 23.C. Zhao, Z. Qu, X. Jiang, Y. Tu and X. Bai, "Content-Adaptive Auto-Occlusion Network for Occluded Person Re-Identification," in IEEE Transactions on Image Processing, vol. 32, pp. 4223-4236, 2023.
3. Wu, T., Zhang, S., Chen, D. and Hu, H. (2023), Multi-level cross-modality learning framework for text-based person re-identification. *Electron. Lett.*, 59: e12975. https://doi.org/10.1049/ell2.12975.
4. Lu, Z., Lin, R., Deng, H., Hu, H. and Chen, Z. (2022), Common visual part alignment for vehicle re-identification. *Electron. Lett.*, 58: 399-401. https://doi.org/10.1049/ell2.12457.
5. T. Wang, H. Liu, P. Song, T. Guo, and W. Shi, "Pose-guided feature disentangling for occluded person re-identification based on transformer," in *Proc. AAAI Conf. Artif. Intell.*, 2022, pp. 2540–2549.
6. S. Kim, S. Kang, H. Choi, S. S. Kim and K. Seo, "Keypoint Aware Robust Representation for Transformer-Based Re-Identification of Occluded Person," in *IEEE Signal Processing Letters*, vol. 30, pp. 65-69, 2023.
7. S. Dou, C. Zhao, X. Jiang, S. Zhang, W. -S. Zheng and W. Zuo, "Human Co-Parsing Guided Alignment for Occluded Person Re-Identification,"in *IEEE Transactions on Image Processing*, vol. 32, pp. 458-470, 2023.
8. P. Wang, C. Ding, Z. Shao, Z. Hong, S. Zhang and D. Tao, "Quality-Aware Part Models for Occluded Person Re-Identification," in *IEEE Transactions on Multimedia*, vol. 25, pp. 3154-3165, 2023.
9. Z. Wang, F. Zhu, S. Tang, R. Zhao, L. He and J. Song, "Feature Erasing and Diffusion Network for Occluded Person Re-Identification," *2022 IEEE/ CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, LA, USA, 2022, pp. 4744-4753.
10. B. Xu, L. He, J. Liang and Z. Sun, "Learning Feature Recovery Transformer for Occluded Person Re-Identification," in *IEEE Transactions on Image Processing*, vol. 31, pp. 4651-4662, 2022.
11. Y. Xu et al., "Vitpose: Simple vision transformer baselines for human pose estimation," *Advances in Neural Information Processing Systems* 35 (2022): 38571-38584.
12. A. V. Oppenheim and J. Lim, "The importance of phase in signals," *Proc. IEEE*, vol. 69, no. 5, pp. 529–541, May 1981.
13. X. He, Z. Mo, P. Wang, Y. Liu, M. Yang and J. Cheng, "ODE-Inspired Network Design for Single Image Super-Resolution," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 1732-1741.
14. J. Anderson. Computational fluid dynamics: the basics with applications. Multidisciplinary Digital Publishing Institute, 1995. 4.
15. A. Hermans, L. Beyer, and B. Leibe. In Defense of the Triplet Loss for Person Re-Identification. 3 2017.
16. J. Xue et al., "Clothing Change Aware Person Identification," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Salt Lake City, UT, USA, 2018, pp. 2193-21938.
17. J. Miao, Y. Wu, P. Liu, Y. Ding, and Y. Yang, "Pose-guided feature alignment for occluded person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 542–551.
18. J. Zhuo, Z. Chen, J. Lai, and G. Wang, "Occluded person re-identification," in *Proc. IEEE Int. Conf. Multimedia Expo.*, 2018, pp. 1–6.
19. L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1116–1124.
20. Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled samples generated by GAN improve the person re-identification baseline in vitro," in *Proc. IEEE Int. Conf.Comput. Vis.*, 2017, pp. 3754–3762.
21. K. Zhou, and T. Xiang. (2019). "Torchreid: A Library for Deep Learning Person Re-Identification in Pytorch," ArXiv, abs/1910.10093.
22. Y. Li, J. He, T. Zhang, X. Liu, Y. Zhang and F. Wu, "Diverse Part Discovery: Occluded Person Re-identification with Part-Aware Transformer," *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Nashville, TN, USA, 2021, pp. 2897-2906.