Self-guided Body Part Alignment with Relation Transformers for Occluded Person Re-Identification

Guanshuo Wang, Xiong Chen, Jialin Gao, Xi Zhou, Shiming Ge, *Senior Member, IEEE*

***Abstract*—Person re-identification in the wild is often challenged by occlusion. Existing methods mainly rely on learned external cues like pose or parsing to ease occlusion distraction. This knowledge highly related to body semantics may introduce alignment effects, leading to additional requirements for dedicated training data and inference computation. We propose the Self-guided Body Part Alignment method that learns cue-free semantic-aligned local prediction for feature representations to avoid high-cost dependence on external cues. First, scale-wise global spatial attention is utilized to determine essential body parts automatically. A relation transformer network is then employed to predict semantic-aligned local parts, guided with anchored global information by constraint loss. Similarity metrics for all parts are merged with threshold conditions to filter invisible body parts comprehensively. Experimental results on occluded and holistic person reID benchmarks show the proposed method outperforms other cue-relied and cue-free methods. As far as we know, this is the first method that applies transformer networks on local predictions for occluded reID tasks.**

***Index Terms*—Deep learning, person re-identification, transformer network, attention mechanism**

I. Introduction

Person re-identification (reID) aims to retrieve pedestrian images from multi-view camera captures to locate or track a specific identity. Recent person reID approaches base on deep learning focuses on extracting discriminative features from part-aligned holistic pedestrian images. However, this body part alignment assumption is far from practice. On the one hand, the detected pedestrians are often occluded by foreground objects or other pedestrian bodies, leading to occluded person reID issues. On the other hand, some parts from probe pedestrian bodies are often excluded from bounding boxes due to incomplete detection or capture, called partial person reID problems. Both these incomplete body situations can break the alignment consistency of training domain distribution and distract the effectiveness of discriminative representation for accurate retrieval. Therefore, it is practical to address incomplete reID issues for performance improvement.

To this end, occluded person reID are extensively focused by the field, approaches for which are mainly divided into cue-relied and cue-free methods. Cue-relied methods focus on visible parts and supplement body part alignment by external cues from dedicated learning-based methods. Accurate pose estimation [1], [2], [3], [4] provides predefined structural body joints and skeletons information, widely applied for alignment guidance in occluded reID [5], [6], [7], [8]. Human body parsing gives pixel-level part prediction with semantic correspondence [9], [10]. The commonality of these methods is that their relied external cues are all prompts with high computing and data requirements, which can be a bottleneck in practical deployment.

Unlike cue-relied methods, cue-free methods [5], [11], [12] have no explicit guidance by external cues. [5] tends to apply data augmentation to simulate rare occluded pedestrian samples. [11] employ iterative clustering along the spatial dimension of feature maps for unsupervised part parsing for semantic alignment. [12] focuses on the set intersection conflict penalty along the channel dimension patterns by an end-to-end training strategy. However, these methods do not solve the problem very efficiently, and introduce repeated calculations and convergence difficulties.

From the view of supervision, external cues can be regarded as annotated knowledge of the body structure. The mentioned cue-free methods pay attention to data augmentation, unsupervised clustering, and regularization, which are primarily essential aspects in self-supervised learning [13], [14], [15], a paradigm that has made great success in general tasks with insufficient label annotations. Recently attention mechanism has been exploited for self-supervised methods [16], [17], which indicates new ideas in more effective cue-free approaches.

In this letter, we propose the Self-guided Body Part Alignment (SBPA) method. Compared with cues-relied methods, it can achieve even more effective alignment, but avoid heavy inference costs for external semantics guidance. SBPA learns semantic-aligned local prediction to locate significant body parts for more discriminative feature representations. First, the global attention map with cross-scale information is generated upon representations for the body content augmentation and the target for subsequent guidance. Then, semantic-aligned part features determined by independent local predictions are represented with relation message passing by a transformer network. During training, a self-guided constraint is applied to guide the local predictions to be heterogeneous and semantic-aligned by minimizing the difference between local and anchored global attentions. Feature similarities are calculated with conditional metrics upon part visibility scores. Experiments show SBPA can outperform the top cue-relied and cue-free methods on occluded or holistic person reID.

This work was partially supported by grants from the National Natural Science Foundation of China (61772513). (Corresponding author: Shiming Ge, e-mail: geshiming@iie.ac.cn)

G. Wang and J. Gao are with Cooperative Medianet Innovation Center, Shanghai Jiao Tong University, Shanghai, 200240, China (e-mail: guanshuo.wang@sjtu.edu.cn).

X. Chen and X. Zhou are with CloudWalk Technology, Shanghai, 201203, China.

S. Ge is with the Institute of Information Engineering, Chinese Academy of Sciences, Beijing, 100095, China, and is also with School of Cyber Security, University of Chinese Academy of Sciences.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/LSP.2021.3087079, IEEE Signal Processing Letters

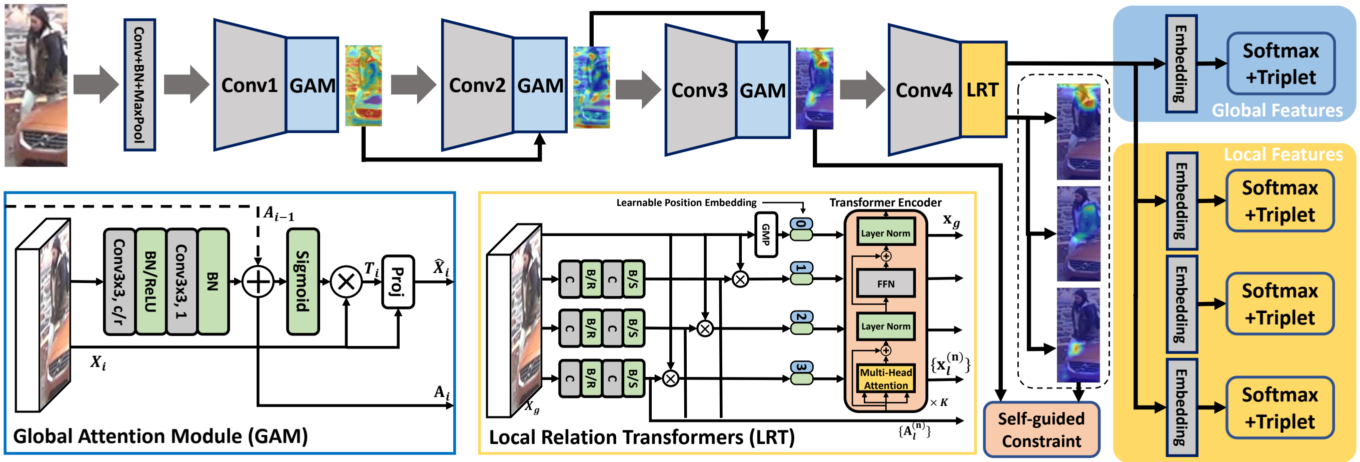


Fig. 1. The framework of Self-guided body Part Alignment. This figure shows the backbone network with our proposed modules, including scale-wise Global Attention Modules (GAM), Local Relation Transformer (LRT), and the Self-guided Constraint Loss (SCL). GAM generates spatial attention maps across different scales in forwarding propagation for feature augmentation. LRT predicts local predictions and feature representations with message passing by Transformer networks. SCL applies correspondence to the generated local attention by predefined spatial stripe anchor attentions.

II. Proposed Approach

Fig. 1 demonstrates the framework of the Self-guided Body Part Alignment method. We look into semantic-aligned salient content mining with the "global before local" strategy. Global Attention Modules and Local Relation Transformer provides the spatial attention hints and establishes the guidance by the self-guided constraint loss.

A. Cross-scale Global Attention

Global spatial attention can be a powerful feature augmentation [18] with important guidance to gather different significant body parts. Considering a balance between performance and computational costs, we design the Global Attention Module (GAM) for learning a spatial attention map with significant part augmentation in different scales. The attention map $\mathbf{A}_i$ for the feature map $\mathbf{X}_i$ in i-th scale is given by

$$\mathbf{A}_i = \begin{cases} f_g(\mathbf{X}_i), & i = 1, \\ f_g(\mathbf{X}_i) + \lfloor \mathbf{A}_{i-1} \rfloor, & i > 1, \end{cases} \tag{1}$$

Here $f_g(\mathbf{X}_i)$ denotes the 2-layer CNN with Conv2d and BatchNorm operations to predict the attention map in i-th scale, and $\lfloor \mathbf{A} \rfloor$ is the downsampled map as coarse guidance to finer scale attentions. For scales smaller than the initial, attention maps are determined by both the response in previous and current scales. All GAMs are employed after residual conv blocks for feature augmentation. Token features $\mathbf{T}_i$ for important semantic abstraction by merging the augmentation with global average pooling, *i.e.* $\mathbf{T}_i = GAP(Sigmoid(A_i) \odot X_i)$. Output feature maps $\hat{\mathbf{X}}_i$ can be refined by bilinear projection as

$$\hat{\mathbf{X}}_i = \mathbf{X}_i + (q(\mathbf{X}_i)^T \cdot k(\mathbf{T}_i)) \times v(\mathbf{T}_i), \tag{2}$$

q, k and v are linear embedding with channel dimension d_i. This projection operation implies the abstract significance token across global spatial dimensions, but brings lighter computation costs than non-local self-attention and avoids numerical issues for more robust convergence than dot product.

B. Local Relation Transformer

As shown in Fig. 1, GAMs in different scales can generate global attention maps to emphasize essential body parts around the pedestrians. However, it is far from semantic-level body part alignment. Here we borrow the architecture of GAM to determine semantic-aligned local predictions. Here we employ N independent CNNs to generate local spatial attention maps, denoted as $\mathbf{A}_l^{(n)} = f_l^{(n)}(\mathbf{X})$. The local tokens are also introduced similarly with GAMs, *i.e.* $\mathbf{T}_l^{(n)} = GAP(A_i \odot X_i)$, which can be regarded as the discriminative local features.

Learning local prediction in these duplicated architecture groups tends to be a homogeneous set without proper message interaction. Recent transformer networks [19] are widely used to deal with sequential data [20] or set prediction problems [21], which might be a hint in our local prediction problem. Based on the basic part spatial attention, we introduce the Local Relation Transformer (LRT) module in local representation learning. The framework is shown in Fig. 1. A transformer encoder layer consists of several operations: 1) The input embedding sequence z_0 is concatenated from both the globally pooled feature and all local tokens $\mathbf{T}_l^{(n)}$ from convolutional local prediction. A learnable position embedding e_p is appended to the input sequence to retain positional information of tokens. 2) Multi-head Self-Attention (MSA) is applied to establish the sequential relationship between local predictions. 3) FFN is a 2-layer fully connected network module with an intermediate GELU activation. 4) Sequences from both MSA and FFN modules are connected with residual addition and applied with layer normalization on the sequential dimension. 5) We can stack K layers of transformers to increase the node-wise connection and non-linearity, with final output embeddings for global $\mathbf{x}_g$ and local $\mathbf{x}_l^{(n)}$ at the top.

C. Self-guided Constraint Loss

Guidance for separating corresponding local semantic spatial attention from global is necessary for our proposed framework. We design a self-guided constraint loss to help Relation

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/LSP.2021.3087079, IEEE Signal Processing Letters

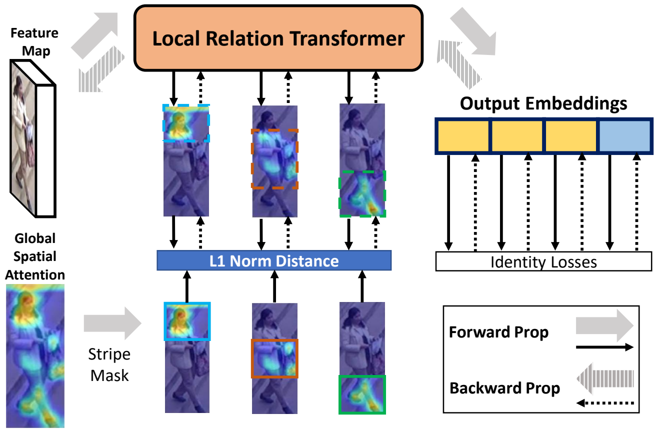


Fig. 2. A simplified pipeline of self-guided constraint upon Local Relation Transformer. The local spatial attention maps with specific semantics are constrained by L1-norm distance to the stripe masked global attention and the identity discrimination on output embedding. The backward propagation to the global spatial attention is isolated.

Transformers to learn local predictions, as shown in Fig. 2. Inspired by stripe-based feature learning, splitting feature maps into horizontal stripe regions can be effective in pooling local features. Here we similarly anchor the last global spatial attention map $\mathbf{A}_g$ into N masked rigid maps $\mathbf{A}_M^{(n)}$ as

$$\mathbf{A}_M^{(n)}(i,:) = \begin{cases} \mathbf{A}_g(i,:), & i \in [\frac{nH}{N}, \frac{(n+1)H}{N}), \\ 0, & i \in others, \end{cases} \tag{3}$$

And here we achieve the guidance with an attention distilling from the masked global to the local predictions by the loss L_p, which minimizes the L1 norm distance between the attention from local prediction with the masked given by

$$L_p = \frac{1}{N} \sum_{n=0}^{N-1} \|\mathbf{A}_l^{(n)} - \mathbf{A}_M^{(n)}\|_1, \tag{4}$$

However, we should notice that simple attention distilling tends to equivalent effects to rigid splitting that has been blamed with inadequate robustness to misalignment, opposite to our motivation for body part alignment. LRT can avoid these effects by node-wise message passing. In Fig. 2, comparing the second and third stripe masks from global attention and the corresponding local prediction attention from LRT, the inference result is more comprehensive to the whole body part with the same semantic. On the other hand, without the part constraint, LRT may also fall into the trap of homogenization.

D. Conditional Metrics

Local predictions with specific semantics can be determined. Subsequently, we need to calculate the feature similarity matrix for retrieval ranking. Previous metrics for occluded reID [22], [7] utilize predicted visibility scores to merge similarities across different parts. Here we use the average part prediction scores of top-$\lfloor HW/N \rfloor$ values to model the visibility v_n of local prediction for part n, calculated as

$$v_n = \frac{1}{\lfloor HW/N \rfloor} \sum_{h,w}^{H,W} top(\mathbf{A}_l^{(n)}(h,w), \lfloor HW/N \rfloor), \tag{5}$$

We define that when v_n is larger than a threshold v_{th}, this part can be regarded as visible. Instead of weighted summation, we take v_n as a condition to switch between weighted and statistic metrics to enhance the effects of salient visible parts in retrieval ranking. The conditional similarity metric $s(q,g)$ between query q and gallery g images can be calculated as

$$s(q,g) = \begin{cases} \frac{\sum_{n=1}^{N} v_n^{(q)} v_n^{(g)} s_n}{\sum_{n=1}^{N} v_n^{(q)} v_n^{(g)}}, & \forall n: \ v_n^{(q)} \geq v_{th} \cap v_n^{(g)} \geq v_{th}, \\ \max\limits_{\substack{n \in \{1..N\} \\ v_n^{(q)} \geq v_{th} \\ v_n^{(g)} \geq v_{th}}} s_n, & \exists n: \ v_n^{(q)} < v_{th} \cup v_n^{(g)} < v_{th}, \end{cases} \tag{6}$$

where the part similarity is denoted as $s_n = sim(h_n^q, h_n^g)$. When all the parts are visible enough, the weight summation metric can comprehensively associate similarities across parts. When some parts are unseen, we choose to amplify the most important visible metric by max pooling, which can reduce distraction from low visibility parts.

III. Experiments

A. Experimental Settings

Input images in any experimental setting are resized to 384 × 128. For comparison fairness, we use a widely used modified ResNet-50 backbone network [23]. The batch size for training sampling is 48, consisting of randomly sampled 12 identities with 4 images per identity. The margin for batch hard triplet loss is set to 0.6. We train the model using SGD optimizer with momentum 0.9 for 60 epochs. The initial learning rate is set to 0.01, decayed by factor 0.1 after 30 and 50 epochs. For occluded tasks, we evaluate the model on the Occluded-DukeMTMC dataset. For the holistic tasks, we evaluate the model on Market-1501 and DukeMTMC-reID datasets. The inference time is about 33ms per batch on our platform, slower than the baseline by about 30%.

B. Comparison Results

Comparison for occluded reID is shown in Table I. SBPA outperforms the top reported method by at least 1.7%/1.7% Rank-1 and mAP. Some surprising conclusions can be obtained from the results: 1) SBPA achieves obvious improvement with the baseline, which confirms effects on our tasks. 2) The performance margin between pose-guided HONet, parsing-guided SGAM, and our self-guided SBPA shows external cues are not essential to conquer occlusion. 3) Compared with the ISP by iterative clustered parsing, SBPA achieves better results relying on alignment with lighter computation. These reveal a great potential of attention on local information decoupling.

Holistic results are shown in Table II. Compared with state-of-the-art attention-based method ABD-net, SBPA outperforms by mAP/Rank-1 on Market-1501 and DukeMTMC-reID. Considering our method does not benefit from channel attention, it shows the effectiveness of local prediction by self-guided alignment in some aspects. MGN utilizes rigid stripe-based aligned features with multiple granularities for more powerful discrimination. SBPA with more flexible local features shows a performance priority on self-guided prediction.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/LSP.2021.3087079, IEEE Signal Processing Letters

TABLE I
PERFORMANCE COMPARISON RESULTS WITH OTHER STATE-OF-THE-ART METHODS ON OCCLUDED-DUKEMTMC DATASET FOR THE OCCLUDED PERSON REID TASK. (†: MODEL WITH HEAVY BACKBONE HRNET-W32; SCL(LP): SELF-GUIDED CONSTRAINT LOSS APPLIED ON MULTI-HEAD LOCAL PREDICTIONS WITH NO TRANSFORMERS)

Method	R-1	R-5	R-10	mAP
SFR [9]	42.3	60.3	67.3	32.0
PGFA [6]	51.4	68.6	74.9	37.3
SGAM [10]	55.1	68.7	74.0	35.3
HONet [8]	55.1	-	-	43.8
Adver Occluded [24]	44.5	-	-	32.2
PCB [25]	42.6	57.1	62.9	33.7
MoS [12]	61.0	-	-	49.2
ISP† [11]	62.8	**78.1**	82.9	52.3
BOT [23] (Baseline)	53.2	67.9	74.0	40.8
BOT+GAM	60.0	76.6	81.2	49.2
BOT+GAM w/o scale-wise	59.3	76.0	80.8	48.9
BOT+GAM w/o bilinear	56.1	72.5	77.9	44.3
BOT+GAM+LRT	62.6	77.2	82.4	51.9
BOT+GAM+SCL(LP)	61.8	77.1	81.7	52.0
BOT+GAM+LRT+SCL	64.0	77.4	82.6	52.8
SBPA (Ours)	**64.5**	78.0	**82.9**	**54.0**

TABLE II
COMPARISON WITH OTHER STATE-OF-THE-ART METHODS FOR HOLISTIC PERSON REID TASKS.

Methods	Market-1501		DukeMTMC-reID	
	mAP	Rank-1	mAP	Rank-1
HA-CNN [26]	75.7	91.2	63.8	80.5
PCB+RPP [25]	81.6	93.8	69.2	83.3
HPM [27]	83.1	93.9	74.5	86.3
HONet [8]	84.9	94.2	75.6	86.9
MoS [12]	86.8	94.7	77.0	88.7
MGN [28]	86.9	95.7	78.4	88.7
ABD-Net [29]	88.3	95.6	78.6	89.0
SBPA w/o CM	88.3	95.7	77.7	88.7
SBPA (Ours)	**89.0**	**96.0**	**78.9**	**89.6**

C. Effectiveness of Components

In Table I, we evaluate the effectiveness of components. GAM contributes the major improvement for the necessity of significant parts by spatial attention, which are also benefited by the scale-wise connection and bilinear projection. Based on this setting, applying LRT or local predictions with SCL respectively bring similar improvement effects, and LRT is more effective on Rank-1. Combining LRT and SCL brings a further improvement, which verifies the alignment effects of relation message passing by LRT and constraint loss. Applying conditional metrics (CM) makes the complete SBPA framework and brings salient improvement, especially on mAP.

D. Parameter Analysis

Evaluation on part number N**.** The part number controls the granularity of local features. Fig. 4(a) shows the line chart about Rank-1 accuracy and mAP along different N. We can find N larger than 3 does not perform well, which might be suffered from too fine-grained stripe anchors that make the guidance hard to discriminate local features in semantic levels.

Evaluation on Transformer depth K**.** The Transformer depth in LRT determines the message passing depth along with the sequential nodes and the extra computation costs for local predictions. Fig. 4(b) shows the line chart along different K. We find that $K = 2$ catches the best performance, and further depth Transformers introduce obvious performance degradation and slow down inference efficiency.

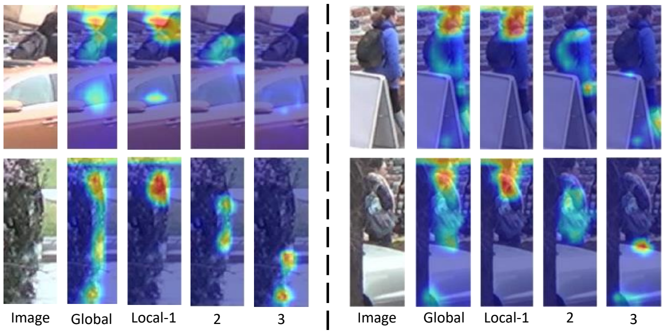
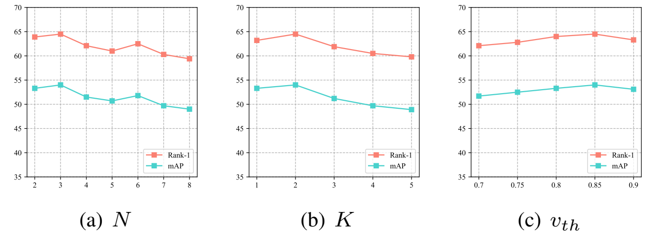


Fig. 3. Visualization for global and local attention maps in SBPA with $N = 3$.



(a) N (b) K (c) v_{th}

Fig. 4. Impact of parameters to Rank-1 and mAP for occluded person reID.

Impact of metric threshold v_{th}**.** This parameter controls the pathway for conditional metrics. Fig. 4(c) demonstrates the relationship between metric accuracy and threshold. We find that both Rank-1 and mAP reach their peak around 0.85, and too loose or tight might affect the ranking performance.

E. Visualization on Part Alignment

Fig. 3 demonstrates several visualization results of global and local spatial attention maps by GAM and LRT in SBPA with $N = 3$. From the global view, we can see the significant body content is highly responsive for further feature augmentation, and non-body occlusion content is mainly excluded from the response even if body parts are severely occluded by thick bush, which provides reliable guidance for local part alignment in occlusion. From the local view, we can see the responsive parts in corresponding semantic attention maps are well aligned, and completely occluded parts like feet or legs are precisely filtered in the results.

IV. CONCLUSION

In this letter, we propose the cue-free Self-guided Body Part Alignment (SBPA) framework for occluded person reID. SBPA can learn local part predictions with consistent semantics mined from spatial attention. Effective global spatial attention and local relation transformers are applied to introduce powerful feature augmentation and relation message passing. The self-guided constraint is imposed for spatial semantic alignment. Efficient conditional metrics are employed on local representations. Extensive experiments on occluded reID datasets show that the proposed method achieves competitive performance compared with the state-of-the-art methods.

References

[1] C. Su, J. Li, S. Zhang, J. Xing, W. Gao, and Q. Tian, "Pose-driven deep convolutional model for person re-identification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2017, pp. 3980–3989.

[2] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu, "RMPE: Regional multi-person pose estimation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2017, pp. 2353–2362.

[3] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1302–1310.

[4] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh, "Openpose: Realtime multi-person 2d pose estimation using part affinity fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019, early Access.

[5] J. Zhuo, Z. Chen, J. Lai, and G. Wang, "Occluded person re-identification," in *The IEEE International Conference on Multimedia and Expo*, 2018, pp. 1–6.

[6] 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/CVF International Conference on Computer Vision*, 2019, pp. 542–551.

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

[8] 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/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 6449–6458.

[9] 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/CVF International Conference on Computer Vision*, 2019, pp. 8450–8459.

[10] Q. Yang, P. Wang, Z. Fang, and Q. Lu, "Focus on the visible regions: Semantic-guided alignment model for occluded person re-identification," *Sensors*, vol. 20, no. 16, p. 4431, 2020.

[11] 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*, 2020, pp. 346–363.

[12] M. Jia, X. Cheng, Y. Zhai, L. Shijian, S. Ma, Y. Tian, and J. Zhang, "Matching on sets: Conquer occluded person re-identification without alignment," in *the AAAI Conference on Artificial Intelligence*, 2021.

[13] M. Caron, P. Bojanowski, A. Joulin, and M. Douze, "Deep clustering for unsupervised learning of visual features," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 132–149.

[14] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9729–9738.

[15] X. Zhan, J. Xie, Z. Liu, Y.-S. Ong, and C. C. Loy, "Online deep clustering for unsupervised representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 6688–6697.

[16] Y. Wang, J. Zhang, M. Kan, S. Shan, and X. Chen, "Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation," in *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 12 272–12 281.

[17] A. Deshpande and K. Narasimhan, "Guiding attention for self-supervised learning with transformers," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 2020, pp. 4676–4686.

[18] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European Conference on Computer Vision*, September 2018.

[19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[20] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, 2019, pp. 4171–4186.

[21] 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*, 2020, pp. 213–229.

[22] 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/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 393–402.

[23] 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/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019.

[24] H. Huang, D. Li, Z. Zhang, X. Chen, and K. Huang, "Adversarially occluded samples for person re-identification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5098–5107.

[25] 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*, 2018, pp. 480–496.

[26] W. Li, X. Zhu, and S. Gong, "Harmonious attention network for person re-identification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2285–2294.

[27] Y. Fu, Y. Wei, Y. Zhou, H. Shi, G. Huang, X. Wang, Z. Yao, and T. Huang, "Horizontal pyramid matching for person re-identification," in *the AAAI Conference on Artificial Intelligence*, 2019, pp. 8295–8302.

[28] G. Wang, Y. Yuan, X. Chen, J. Li, and X. Zhou, "Learning discriminative features with multiple granularities for person re-identification," in *ACM International Conference on Multimedia*, 2018, pp. 274–282.

[29] T. Chen, S. Ding, J. Xie, Y. Yuan, W. Chen, Y. Yang, Z. Ren, and Z. Wang, "Abd-net: Attentive but diverse person re-identification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 8351–8361.