2021 IEEE International Conference on Image Processing (ICIP) | 978-1-6654-4115-5/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICIP42928.2021.9506010

PCNET: PARALLELLY CONQUER THE LARGE VARIANCE OF PERSON RE-IDENTIFICATION

Jianyuan Wang[1], *Meiyue You*[2], *Biao Leng*[3,*], *Ming Jiang*[3], *and Guanglu Song*[3]

[1]School of Automation Science and Electrical Engineering, Beihang University, Beijing, 100191
[2]School of Information Science and Technology, Beijing University of Chemical Technology, Beijing, 100029
[3]School of Computer Science and Engineering, Beihang University, Beijing, 100191

ABSTRACT

Person re-identification has a wide range of applications, and many state-of-the-art methods are proposed to solve the problem under specific scenarios. However, it is still a challenging issue because of the large variance in practical applications, such as pose variations, misalignment, and image noises. In this paper, Parallelly Conquer Net (PCNet) is proposed to deal with large variance in a parallel manner. PCNet consists of three module: Pose Adaptation Module (PAM), Global Alignment Module (GAM), and Pixel-Wised Attention Module (PWAM). Each module is designed to deal with a sub-variance independently. Furthermore, the generated features are aggregated by parallel branches to utilize complementary information among them. Extensive experiments on three benchmarks (Market-1501, DukeMTMC-reID, and CUHK03) demonstrate the effectiveness of the method. The results show that PCNet can significantly improve the performance of person re-identification.

***Index Terms*—** Person Re-identification, Parallelly Conquer Net, Pose Adaptation, Alignment, Pixel Attention

1. INTRODUCTION

Person re-identification (person re-id) [1] has obtained an increasing attention due to its important applications in monitoring systems. As the development of Convolutional Neural Network (CNN), the performance of person re-id has been significantly improved [2, 3]. However, it remains a challenging task because of the large variance. For example in Fig. 1, (a) the poses of the same person tend to be different caused by human movements; (b) misalignment often occurs when images are captured by detectors; (c) the uncontrollable image noises (including occlusions and background clutters) make the person retrieval even more difficult; (d) in some hard cases, images may contain all these variations.

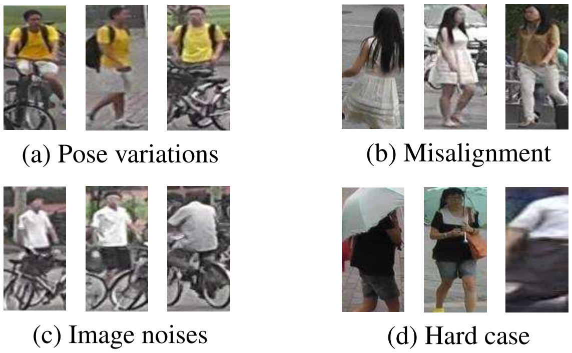
(a) Pose variations
(b) Misalignment
(c) Image noises
(d) Hard case

Fig. 1. Examples in Market-1501 dataset. For each figure, the first is query image, and the second is Rank-1 image of the PCNet, which is the same person. The third is Rank-1 image of the baseline, which is mismatching.

In recent years, deep learning methods are usually used to extract discriminative features [4] for person re-id task. For instance, the image size [5], scale [6], alignment [7], segmentation [8] are taken into account to improve the feature representation. Spatial transformer network (STN) [9] and the partial framework [10] are designed to observe competition between STN module and ReID module. Furthermore, the part based deep learning models have shown superior performance in many literatures [11, 12, 13]. Part based models can help to learn more fine-grained part features which are conducive to identify different persons [14]. Different approaches have been proposed to locate and align human parts across images, which can help handle the part-misalignment problem caused by the variations of viewpoints and poses [15, 16]. However, existing methods mainly focus on eliminating a specific variance [17, 18], which is not robust enough for complex scenarios.

Considering the complex scenarios in person re-id, PCNet is proposed to eliminate the large variance in a parallel manner with three sub-modules. Each module is designed to solve a kind of variance (e.g. pose variations, misalignment, and noises). At last, PCNet integrates features generated by different sub-modules together through the Module Aggregation Unit to further improve the performance and robustness

[*]Corresponding author

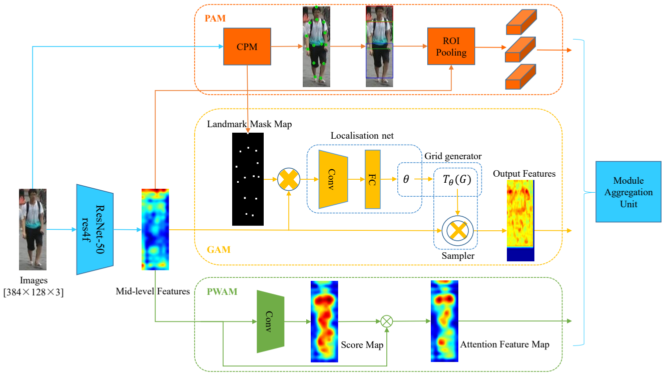

Fig. 2. Illustration of the proposed PCNet. It consists of three parallel sub-modules and a Module Aggregation Unit.

of the proposed method.

To sum up, the main contributions of this paper are as follows: (a) the Pose Adaptation Module (PAM) is embedded in the parallel model to locate the precise body regions. (b) the Global Alignment Module (GAM) is proposed to align the person where the person landmarks are further used to generate the affine transformation matrix. (c) the Pixel-Wised Attention Module (PWAM) adopts attention mechanism to depose the impact of the background and occlusion.

2. PROPOSED METHOD

2.1. Overview of the Parallelly Conquer Net

The framework of PCNet is shown in Fig. 2. In this framework, we disentangle the pose variation, misalignment, and image noises from person re-id and then adopt the parallel sub-modules to solve them. PAM is designed to extract the features of precise body regions via the detected human landmarks. GAM is designed to make the person landmarks and STN [9] more adaptive to person images. PWAM employs an attention mechanism to make the network focus on the human body in images while ignoring image noises. Given an image, PCNet firstly extracts the mid-level features via the fully convolutional network (FCN), then feeds the features into the three parallel sub-modules. Finally, these features are aggregated by the Module Aggregation Unit.

2.2. Modules in Parallelly Conquer Net

2.2.1. Pose Adaptation Module

PAM is designed to alleviate the large pose variations caused by pre-defined division of human body. In this module, the Convolutional Pose Machines (CPM) [19] is adopted to detect fourteen human body landmarks, and then to get the bounding boxes to divide the human bodies into different regions. Let $\mathcal{S}_i (i = 1, 2, 3)$ denotes the human body landmarks set of the ith region, then $\mathcal{S}_1 = \{1, 2, 3, 6\}$, $\mathcal{S}_2 = \{3, 4, 5, 6, 7, 8, 9, 12\}$, $\mathcal{S}_3 = \{9, 10, 11, 12, 13, 14\}$ according to the index number in Fig. 2. The bounding box of the ith region $\mathcal{B}_i (i = 1, 2, 3)$ can be formulated as:

$$\mathcal{B}_i = [x_{i1}, y_{i1}, x_{i2}, y_{i2}] = [1, \min_{j \in \mathcal{S}_i}(y_j), W, \max_{j \in \mathcal{S}_i}(y_j)]. \qquad (1)$$

Where W is the width of the image. $(x_{i,1}, y_{i,1})$ and $(x_{i,2}, y_{i,2})$ are the upper left and lower right coordinates of the bounding box, respectively. For weakening the impact of biased landmarks, we extend it horizontally to the image boundary.

2.2.2. Global Alignment Module

GAM is proposed to fix misalignment in automatic detected human. In GAM the human landmarks are cooperated with STN to learn a more robust transformation matrix to align the images.

The original STN has three components: a localisation network, a grid generator, and a sampler. In this paper, the localisation network takes the input feature maps and outputs a 2×3 2D affine transformation matrix θ. The grid generator uses θ to create a regular sampling grid, then the sampler produces the sampled output feature maps from the input feature maps using the sampling grid. As for person re-id task, the person images always do not need to be rotated. So we mainly learn four parameters for translation and scale.

Human body landmarks are adopted to guide the learning. For generating the landmark mask map, suppose $L = \{(x_i, y_i)\} (i = 1, 2, 3, \ldots, 14)$ representing the human body landmarks set, we generate the mask $M \in \mathbb{R}^{H \times W}$ as follows:

$$M(x, y) = \begin{cases} 1, & |x - x_i| \leq l \,\&\, |y - y_i| \leq l \\ 0, & otherwise \end{cases} \qquad (2)$$

Where $(x_i, y_i) \in L$. We set $l = 1$ in our experiments. Then the input feature maps $F_I \in \mathbb{R}^{C \times H \times W}$ of the localisation network in GAM can be formulated as:

$$F_I(c, x, y) = F(c, x, y) \cdot M(x, y), \; c \in \{1, 2, ..., C\}. \qquad (3)$$

Where $F \in \mathbb{R}^{C \times H \times W}$ is the mid-level feature maps. With the help of landmark mask, the localisation network can concentrate on the human body more accurately and ignore the impact of other noises without the interference of other redundant image information, .

2.2.3. Pixel-Wised Attention Module

PWAM is designed to alleviate the impact of image noises such as occlusions or background clutters. Fed with the mid-level feature maps, PWAM can predict a probability map where the score in each pixel indicates the importance of the current coordinate. PWAM is a lightweight architecture containing a 1×1 convolutional layer with 512 channels, a 3×3 convolutional layer with 512 channels, and a 1×1 convolutional layer with 1 channel. The activation function of the first two convolutional layers is ReLU. More formally, given the mid-level feature maps $F \in \mathbb{R}^{C\times H\times W}$ as the input of the PWAM, the generated probability map can be formulated as:

$$S = \sigma(f_{AGM}(F)) \in [0, 1], \quad (4)$$

Where $\sigma(\cdot)$ is Sigmoid function and the dimension of S is $H \times W$. Then the corresponding attention feature map F_{att} can be generated as follows:

$$F_{att}(c, x, y) = F(c, x, y) \cdot S(x, y),\ c \in \{1, 2, ..., C\}. \quad (5)$$

Where (x, y) ranges over all spatial positions.

2.3. Module Aggregation Unit

In PCNet, the large variance is disentangled and fed into three parallel sub-modules to extract calibrated features. Considering the complementary information containing in three sub-modules, we propose the Module Aggregation Unit to aggregate them. The final feature of the input image can be written as:

$$F_{final} = [\alpha_1 f_{PAM}, \alpha_2 f_{GAM}, \alpha_3 f_{PWAM}], \quad (6)$$

Where f_{PAM}, f_{GAM}, f_{PWAM} are the features extracted by PAM, GAM, and PWAM, respectively. $\alpha_i\ (i = 1, 2, 3)$ indicates the weights. Suppose a_0 is the Rank-1 accuracy of the baseline, a_1, a_2, a_3 are the Rank-1 accuracies of the three sub-modules in the validation set, respectively. Then we define:

$$\alpha_i = \sqrt{\frac{a_i - a_0}{\sum_{j=1}^{3}(a_j - a_0)}}\ (i = 1, 2, 3). \quad (7)$$

3. EXPERIMENTS

3.1. Datasets and Settings

The experiments are conducted on Market-1501, DukeMTMC-reID and CUHK03, and follow the popular training/testing protocols [17, 20]. These datasets are challenging due to large variance (e.g. pose variations, occlusions and misalignment). Following the standard evaluation protocol, the models are evaluated according to Cumulative Matching Characteristics (CMC) at Rank-1 and mean Average Precision (mAP) on all the three datasets. All the experiments on the three datasets are conducted with single-query setting.

3.2. Implementation Details

The baseline is based on the IDE model [21], which adopts ResNet-50 as the backbone network, and initialized using the ImageNet pre-trained model. During training, we append a Batch Normalization layer after the "pool5" layer, learn features through the combination of triplet loss, center loss and cross entropy loss of ID prediction logits. During testing, we use the output of the "pool5" layer as the feature representations of the images and compute the cosine distances of different features after l^2-normalization. All of our methods are implemented by Caffe.

Table 1. Ablation study results on CUHK03. Rank-1 Accuracy (%) and mAP (%) Are Shown.

Models	detected		labeled	
	Rank-1	mAP	Rank-1	mAP
Baseline	51.36	45.83	55.07	50.16
+PAM/PreS	58.64	52.79	62.14	57.45
+PAM	59.50	53.61	64.50	58.82
+GAM/Fixθ	54.86	48.89	58.07	52.62
+GAM/w/o Mask	55.43	50.22	59.43	54.58
+GAM	56.00	50.66	60.43	55.11
+PWAM/FixS	58.00	52.43	61.29	57.06
+PWAM	58.64	53.07	61.86	57.13
PCNet	63.71	58.54	66.79	63.20

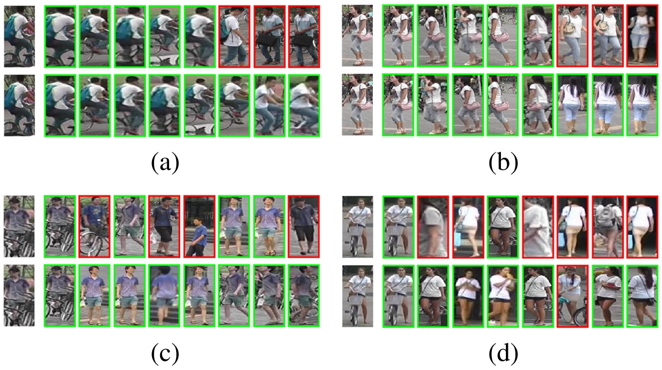

Fig. 3. Examples of the person re-id results on Market-1501 dataset. In each example, The images in the first column are queries, the first row and the second row denote the results of the baseline and the PCNet, respectively. The true positives are in green rectangles while the false positives are in red rectangles.

3.3. Ablation Study

In this subsection, we conduct a fair self-evaluation on CUHK03 dataset to verify the effectiveness of each component of the proposed network. We perform several groups of experiments under the same settings. The results are summarized in Table 1.

Effectiveness of PAM To evaluate the effectiveness of the PAM, the experiments is compared with pre-defined stripes (PAM/PreS), and other structure remains the same. The re-

Table 2. Comparison with State-of-the-art Methods on Market-1501 and DukeMTMC-reID. The Best Results Are Shown in Boldface.

Methods	Market-1501		DukeMTMC-reID	
	Rank-1	mAP	Rank-1	mAP
TriNet+RK [22]	86.7	81.1	–	–
SVDNet+Era [23, 24]	87.1	71.3	79.3	62.4
OGNet [25]	87.7	69.5	76.7	57.9
PAN [17]	88.6	81.5	75.9	66.7
DPFL [5]	88.9	73.1	79.2	60.6
SSP-ReID [20]	92.5	80.1	81.8	68.6
Baseline	90.5	75.8	83.8	66.9
PCNet	**94.6**	**82.8**	**86.4**	**72.7**
+re-ranking	**95.7**	**92.8**	**89.2**	**86.2**

sults are presented in the first two rows in Table 1. When using pre-defined stripes, the Rank-1 accuracy is reduced by 0.86% and 2.36% in the detected dataset and the labeled dataset, respectively. This may be caused by the wrong location of human parts. Thus, the PAM can capture the pose variations better when using human body landmarks.

Effectiveness of GAM GAM/Fixθ in the third row of Table 1 means that the transformation matrix θ is fixed, and do not perform any transformation to the input feature maps. Baseline+GAM/w/o Mask means without the landmark mask, which is similar to [17]. It is shown that STN can help to improve the performance in both detected and labeled dataset. With the help of the landmark mask, the performance is further improved.

Effectiveness of PWAM In order to verify the effect of PWAM, we fix the score maps (PWAM/FixS) to be one. The performance improves by 0.64% and 0.57% in Rank-1 accuracy when the score maps are not fixed. This shows that the PWAM can learn useful score maps from the input feature maps and thus helps to alleviate image noises.

Effectiveness of PCNet The last row in Table 1 presents the results of PCNet. Compared to the results of a single sub-module, the results of the PCNet also have been improved. The result shows that the features learned by these sub-modules are complementary to each other, and the Module Aggregation Unit can further improve the performance.

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

The performance of PCNet is reported in Table 2-3. Compared to the baseline model in this paper, the mAP has been improved by a large margin on all these datasets, which indicates the PCNet has the ability to find more true positives. We also test the results of PCNet combined with re-ranking [27] on Market-1501 and DukeMTMC-reID. Fig.3 illustrates some examples of the person re-id results produced by the baseline and the PCNet.

Table 3. Comparison with State-of-the-art Methods on CUHK03. The Best Results Are Shown in Boldface.

Methods	detected		labeled	
	Rank-1	mAP	Rank-1	mAP
HA-CNN [26]	41.7	38.6	44.4	41.0
PAN [17]	41.9	43.8	43.9	45.8
SVDNet+Era [23, 24]	48.7	43.5	49.4	45.1
TriNet+Era [22, 24]	55.5	50.7	58.1	53.8
Baseline	51.4	45.8	55.1	50.2
PCNet	**63.7**	**58.5**	**66.8**	**63.2**

Comparisons on Market-1501 Compared with baseline, the Rank-1 accuracy and mAP are improved by 4.1% and 7%, respectively. And we increase the Rank-1 accuracy and mAP by 2.1% and 2.7% than the previous best result reported by SSP-ReID [20]. The results demonstrate the effectiveness of the proposed framework on large scale dataset.

Comparisons on DukeMTMC-reID The proposed method also works well on the more challenging dataset. The Rank-1 accuracy and mAP have been improved by 4.6% and 4.1% than SSP-ReID [20]. Compared with baseline, they increased by 2.6% and 5.8%, respectively.

Comparisons on CUHK03 With a small amount of training data, PCNet reaches the best performance. The detected dataset is more challenging than the labeled one, PCNet also exceeds the compared methods by 8.2% and 7.8% in terms of Rank-1 accuracy and mAP, respectively.

4. CONCLUSION

This paper proposes a Parallelly Conquer Net with three well-designed parallel sub-modules to disentangle large variance for person re-id task. The three modules, including Pose Adaptation Module, Global Alignment Module, and Pixel-Wised Attention Module, are employed to deal with pose variations, misalignment and image noises (including occlusions and background clutter), respectively. The experimental results demonstrate the effectiveness and generalization of the proposed method. Each module can improve the accuracy to a certain extent, and Aggregation Unit in the PCNet can well integrate the features of the three sub-modules.

5. ACKNOWLEDGEMENTS

This work is supported by the National Key R&D Program of China (No.2019YFB2102400), the National Natural Science Foundation of China (No.61972014), and the Beijing Municipal Natural Science Foundation (No.L182014) .

6. REFERENCES

[1] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. CH Hoi, "Deep learning for person re-identification: A survey and outlook," *arXiv preprint arXiv:2001.04193*, 2020.

[2] Z. Feng, J. Lai, and X. Xie, "Learning modality-specific representations for visible-infrared person re-identification," *IEEE*, vol. 29, pp. 579–590, 2020.

[3] H. Luo, Y. Gu, X. Liao, S. Lai, and W. Jiang, "Bag of tricks and a strong baseline for deep person re-identification," in *CVPRW*, 2019, pp. 1487–1495.

[4] Y. Xu, Z. Jiang, A. Men, J. Pei, G. Ju, and B. Yang, "Attentional part-based network for person re-identification," in *VCIP*, 2019, pp. 1–4.

[5] Y. Chen, X. Zhu, and S. Gong, "Person re-identification by deep learning multi-scale representations," in *ICCVW*, 2017.

[6] Y. Zhang, W. Shi, S. Liu, J. Bao, and Y. Wei, "Scale-invariant siamese network for person re-identification," in *ICIP*, 2020, pp. 2436–2440.

[7] Y. Wu and W. Sun, "Pedestrian-aligned multiscale features network for person re-identification," in *CAC*, 2019, pp. 362–366.

[8] J. Huang, B. Liu, and L. Fu, "Joint multi-scale discrimination and region segmentation for person re-id," *Pattern Recognit. Lett.*, vol. 138, pp. 540–547, 2020.

[9] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," pp. 2017–2025, 2015.

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

[11] H. Huang, W. Yang, J. Lin, G. Huang, J. Xu, G. Wang, X. Chen, and K. Huang, "Improve person re-identification with part awareness learning," *IEEE Transactions on Image Processing*, vol. 29, pp. 7468–7481, 2020.

[12] X. Lin, Y. Yang, and Z. Niu, "Enhance part-based model for person re-identification with fused multi-scale features," in *ICASSP*, 2020, pp. 4092–4096.

[13] W. Zhang, L. Huang, Z. Wei, and J. Nie, "Appearance feature enhancement for person re-identification," *Expert Systems with Applications*, vol. 163, pp. 113771, 2021.

[14] S. Gao, J. Wang, H. Lu, and Z. Liu, "Pose-guided visible part matching for occluded person reid," in *CVPR*, 2020, pp. 11741–11749.

[15] L. Wei, S. Zhang, H. Yao, W. Gao, and Q. Tian, "GLAD: global-local-alignment descriptor for pedestrian retrieval," in *Proceedings of the 25th ACM International Conference on Multimedia*, 2017, pp. 420–428.

[16] 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 *ECCV*, 2018, pp. 480–496.

[17] Z. Zheng, L. Zheng, and Y. Yang, "Pedestrian alignment network for large-scale person re-identification," *IEEE*, vol. 29, no. 10, pp. 3037–3045, 2018.

[18] K. Han, J. Guo, C. Zhang, and M. Zhu, "Attribute-aware attention model for fine-grained representation learning," in *Proceedings of the 26th ACM international conference on Multimedia*, 2018, pp. 2040–2048.

[19] W. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in *CVPR*, June 2016.

[20] R. Quispe and H. 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.

[21] L. Zheng, Y. Yang, and A. G. Hauptmann, "Person re-identification: Past, present and future," 2016.

[22] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," *arXiv preprint arXiv:1703.07737*, 2017.

[23] Y. Sun, L. Zheng, W. Deng, and S. Wang, "Svdnet for pedestrian retrieval," in *ICCV*, Oct 2017.

[24] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," in *AAAI*, 2020, vol. 34, pp. 13001–13008.

[25] Z. Zheng and Y. Yang, "Parameter-efficient person re-identification in the 3d space," *arXiv:2006.04569*, 2020.

[26] W. Li, X. Zhu, and S. Gong, "Harmonious attention network for person re-identification," in *CVPR*, June 2018.

[27] Z. Zhong, L. Zheng, D. Cao, and S. Li, "Re-ranking person re-identification with k-reciprocal encoding," in *CVPR*, 2017, pp. 3652–3661.