# CROSS-VIEW IDENTICAL PART AREA ALIGNMENT FOR PERSON RE-IDENTIFICATION

*Dongshu Xu*[1,2,3], *Jun Chen*[1,2,3,†], *Chao Liang*[1,2,3], *Zheng Wang*[4], *Ruimin Hu*[1,2,3]

[1]National Engineering Research Center for Multimedia Software, School of Computer Science, Wuhan University, China
[2]Hubei Key Laboratory of Multimedia and Network Communication Engineering, Wuhan University, China
[3]Collaborative Innovation Center of Geospatial Technology, China
[4]National Institute of Informatics, Japan

## ABSTRACT

Person re-identification aims to associate images captured by non-overlapping cameras. It is a challenging task because images are often in different conditions such as background clutter, illumination variation, viewpoint changes and different camera settings. Viewpoint changes and pose variations often cause body part self-occlusion and misalignment. To deal with the problem, local features from human body parts are extracted. However, with viewpoint changes, the body parts also rotate horizontally. It is inappropriate to extract feature from entire area of body parts directly because the visible surface of body parts would turn away if viewpoint changes. Comparing identical areas provides a new way to pay attention to the details of person images. In this paper, we propose a Rotation Invariant Network to find the identical areas in cross-view images to extract robust local features. Extensive experiment show the effectiveness of our method on public datasets including CUHK03, Market1501 and DukeMTMC.

***Index Terms***— Person re-identification, rotation invariant network, identical area comparison

## 1. INTRODUCTION

Person re-identification [1] aims to recognize the same person in non-overlapping cameras. Given a query image containing a person-of-interest and a set of gallary images, it is expected to rank images in the gallary set with visual similarity [2]. In recent years, person re-identification has many important applications in security and video surveilance. For example, search missing persons in shopping center, retrieval suspect in large amount of surveilance videos [3], *etc*.

†Jun Chen is the corresponding author.

This research was supported partially by National Key R&D Program of China (2017YFC0803700), National Nature Science Foundation of China (U1611461, U1736206, 61876135, 61872362, 61671336, 61801335), Technology Research Program of Ministry of Public Security (2016JSYJA12), Hubei Province Technological Innovation Major Project (2016AAA015, 2017AAA123, 2018AAA062), Nature Science Foundation of Hubei Province (2018CFA024) and Nature Science Foundation of Jiangsu Province (BK20160386).

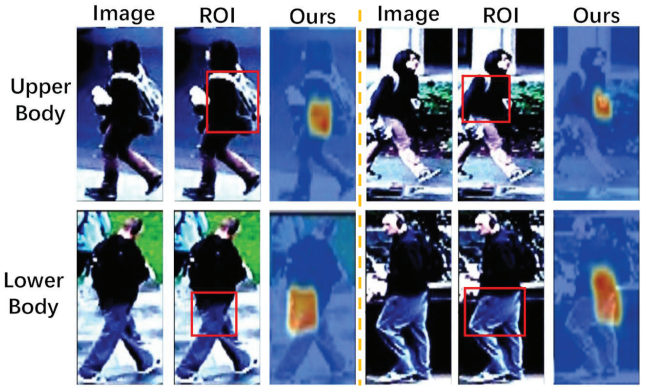


**Fig. 1**. The left images are in one camera, and the right images are in another camera. Our network generates identical part areas of person images in two different cameras while ROIs include entire part area and part background.

Many research works have been proposed to improve performance on public datasets [4][5][6]. However, identifying the same person across different camera views is still a challenging task [7]. Pedestrian images captured by two different cameras often suffer from pose variations which cause a huge intra-class variation in learned visual representation.

Viewpoint changes and pose variations cause body part self-occlusion and misalignment. Previous works utilized horizontal stripes to cope with viewpoint changes in a statistical manner but hard to cope with pose variation. Due to the progress in human pose estimation [9] [10], recent works utilized pose estimation results to align body parts. Specifically, Zheng *et al.* [11] utilized a pose estimation model to obtain keypoints of persons, fixed them by affinity transformation and then compared body part. This method solved the part misalignment problem well but ignoring the occlusion problem of body parts. Zhao *et al.* [12] used a region proposal network, trained on an auxiliary pose dataset, to detect body parts. The methods above extract local features from whole blocks or body parts, where self-occlusion caused by viewpoint changes often leads to information loss. For example in Fig. 1, ROI based method extracts features from entire

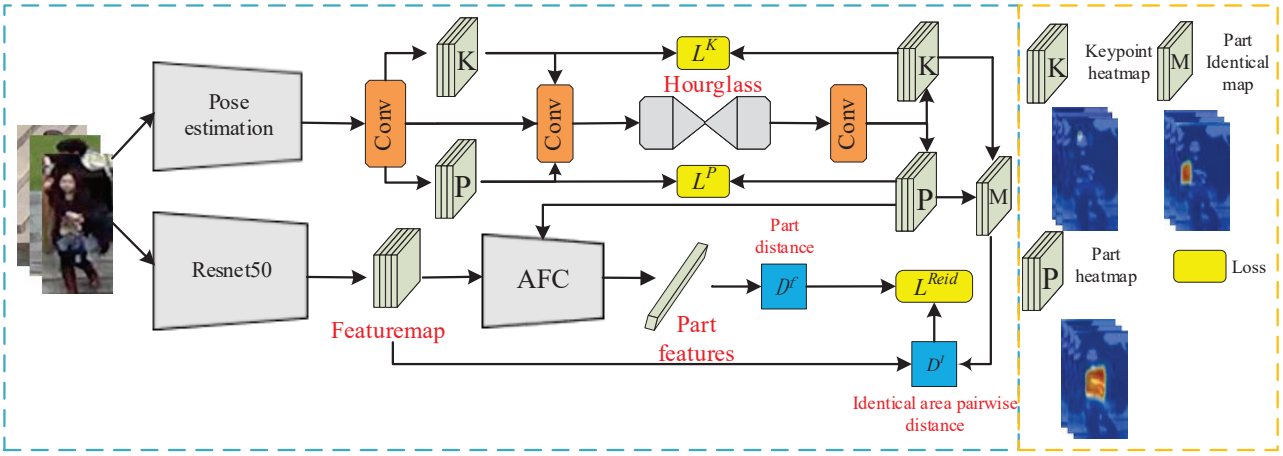


**Fig. 2**. The framework of Rotation Invariant Network (RIN). The upper part is Pose-guided Part Generation Network (PPGN) which generates keypoints and part heatmaps. The lower part is Identical Area Comparison Network which computes distances by part features generated by AFC [8] and identical areas pairwise. The right part illustrates the heatmaps generated by PPGN.

part areas. In the first row in Fig. 1, ROI in the left image contains the right body part with a backpack which is occluded in the right image. In the second row in Fig. 1, ROI in the right image include left leg which is occluded in the left image. Hence, directly extracting features from entire area of body parts including upper body and lower body that often varies with viewpoint changes can deteriorate the matching accuracy. To deal with these problems, finding identical part areas of both stable and dicriminative is important, so that part features can be extracted in same conditions, to alleviate cross-view changes.

In contrast to the works above, we propose to employ keypoints to obtain identical human part areas rather than the entire areas. We argue that comparing identical part areas is naturally more suitable to cope with person re-identification challenges because the clothes are not all in same color or same texture.

In this work, we propose a novel combination network to compare identical part areas in different views. At first, we use a two-stage hourglass based network to acquire person keypoints and part heatmaps. Then we utilize an identical area generation model to obtain intersection area pairwise from the person keypoints and part heatmaps. Finally, we extract features from other parts and identical part areas. Experimental results show effectiveness of the identical area generation mechanism.

## 2. METHODOLOGY

The framework of Rotation Invariant Network (RIN) is illustrated in Fig. 2. RIN consists of two main parts: 1) Pose-guided Part Generation Network (PPGN) and Rotation Invariant Distance Measure (RIDM). The proposed PPGN module aims at generating keypoints by a base network (Hourglass [9] used in this work) and body part masks in two steps. Given a batch of images, PPGN first generates coarse heatmaps of keypoints and body parts, and then refines them. The RIDM obtains keypoints and weighted part masks from PPGN and feature maps from a pretrained re-id base network (Resnet50 [13] fine tuned on re-id datasets). Keypoints can be utilized to generate intersection masks in upper body and lower body which contains most discriminative information. Then AFC [8], a part feature extraction structure, combines weighted part masks and feature maps to acquire part features. The intersection mask and feature maps compose the pairwise intersection features. Finally, a triplet-loss [14] is utilized to learn robust features.

### 2.1. Rotation Invariant Network

Inspired by [9], we utilize two hourglass blocks to construct PPGN by a balance on amount of parameters and prediction accuracy. We take the original hourglass network as a base network, then link a convolutional layer to increase reception field and channels to acquire sufficient part information that part heatmaps demand. After that, we repeat the hourglass structure to make the network stable. When PPGN is fine-tuned in joint network, modules of first hourglass can be fixed so that the second hourglass can learn to generate weighted part heatmap. The labels of part heatmaps are the same as the form of [8].

The PPGN is trained in two steps. In stage 1, we change the kernel in the last convolutional layer from $1 \times 1$ to $3 \times 3$, double the channels and change the output to keypoints. In stage 2, we repeat the structure of stage 1 and change the kernel in input convolutional layer from $1 \times 1$ to $3 \times 3$, and fix the parameters in stage 1. The output of stage $i$ is keypoint heatmaps $K^i$ and part heatmaps $P^i$. The ground truths of

them are $K^*$ and $P^*$. The loss of PPGN is:

$$L^G = L^K + wL^P = \sum_{s=1,2} \|K^s - K^*\| + w\|P^s - P^*\| \quad (1)$$

Based on the selection of identical part area, we utilize AFC [8] to generate part feature so as to combine with our pairwise area distance. In this part, we utilize Resnet50 [13] as the base network to generate feature maps from layer2. Then the feature maps have two ways. One is sent to AFC network to combine with the part masks to generate part features $f$. The other is combined with the intersection masks to generate pairwise distances $D^I$ which is interpreted in Section 2.2. Then we combine pairwise distance of part features $D^f$ and the intersection distance to get the final distance $D$.

$$D = D^f + \lambda D^I \quad (2)$$

Finally, we use triplet-loss [14] as the re-id loss to train the whole network according to the distance matrix $D$.

$$L^{Reid} = \sum_{i}^{N} [D^2_{P_i} - D^2_{N_i} + \alpha]_+ \quad (3)$$

where $D_{P_i}$ is the distance of a positive pair and $D_{N_i}$ is the distance of a negative pair.

### 2.2. Identical area comparison

Viewpoint changes would cause self-occlusion and local area misalignment due to the horizontal rotation of main body parts. It is inapproprate to compare whole body parts directly. We need to find identical areas in different views.

Due to invariance of keypoints in viewpoint changes, we use keypoints [9] which are shown in Fig. 3 to generate identical part areas of the same person in diferent views. Assume that cameras are far away from persons. Given images in camera $C1$ and $C2$, the right image in Fig. 3 shows an illustration of identical area. The ellipse is the transverse plane of upper body. $P1$ and $P2$ are keypoints of left shoulder, which can be viewed as any points in the ecllipse due to viewpoint variations. Define that the parameter equation of the ellipse is:

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1 \quad (4)$$

where $a$ equals a half of upper body height and $b$ is related to waistline and bust which is initialized with a third of $a$.

The $P1$ and $P2$ have the coordinates $(acos\theta_1, bsin\theta_1)$ and $(acos\theta_2, bsin\theta_2)$ seperately. So tangent lines $\overline{L2R2}$ and $\overline{R1R2}$ can be represented as:

$$\frac{xcos\theta_1}{a} + \frac{ysin\theta_1}{b} = 1 \quad (5)$$

$$\frac{xcos\theta_2}{a} + \frac{ysin\theta_2}{b} = 1 \quad (6)$$

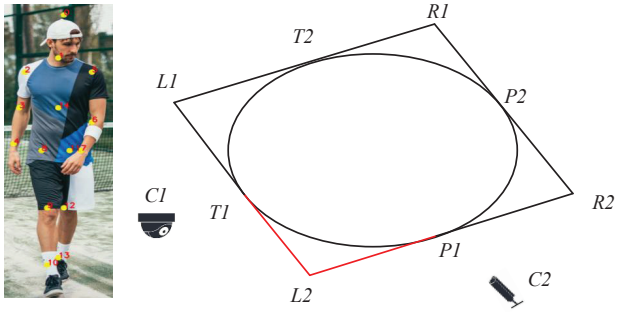


**Fig. 3**. Illustration of keypoints (yellow point) and the intersection area (red line). Utilize the view of transverse plane to deal with horizontal rotation problem.

Combine the two equations above, we can get the horizontal ordinate of the intersection point $R2$:

$$\frac{a(sin\theta_2 - sin\theta_1)}{sin(\theta_2 - \theta_1)} \quad (7)$$

Similarly, the horizontal ordinate of the intersection point $L2$ is formed as:

$$-\frac{a(sin\theta_1 + sin\theta_2)}{sin(\theta_2 - \theta_1)} \quad (8)$$

So the scale factor $\beta$ of the intersection area is:

$$\beta = \frac{|\overline{L2T1}|}{|\overline{L2L1}|} = \frac{sin\theta_2 - sin\theta_1 - cos\theta_2 sin(\theta_2 - \theta_1)}{2sin\theta_2} \quad (9)$$

The distance between two shoulder keypoints in an image is equal to the distance of two parallel tangent lines.

$$d_i = \frac{2}{\sqrt{\frac{cos^2\theta_i}{a^2} + \frac{sin^2\theta_i}{b^2}}} \quad (10)$$

The locations of keypoints determine which quadrant the $\theta_i$ in as well as the signs of $cos\theta_i$ and $sin\theta_i$. If the vertical ordinate of left shoulder keypoint is higher than that of right shoulder, $\theta_i$ is in the first or second quadrant. If the width of left shoulder is higher than that of right shoulder, $\theta_i$ is in the second or third quadrant.

Then we can compute identical area direction by vectors.

$$\overrightarrow{T1P2} = (2acos\theta_1, 2bsin\theta_1) \quad (11)$$

$$\overrightarrow{T2P1} = (2acos\theta_2, 2bsin\theta_2) \quad (12)$$

The clockwise vertical vector to $\overrightarrow{T1P2}$ is:

$$\overrightarrow{n} = (bsin\theta_1, -acos\theta_1) \quad (13)$$

If $\overrightarrow{n} \bullet \overrightarrow{T2P1} > 0$, the scale is $\beta$ and direction is the same as $\overrightarrow{n}$ else $1 - \beta$ and the opposite direction. So the heatmap of identical area can be obtained by part heatmap and the mask which is a vertical stripe, based on the aforementioned scale, direction and keypoint coordinates.

Finally, the pairwise distance $D^I$ is calculated by shortest path distance [15] of identical area heatmaps.

**Table 1**. Comparison results on CUHK03(labeled).

| CUHK03 (labeled) | R-1 | R-5 | R-10 | R-20 |
|---|---|---|---|---|
| SVDNet [19] | 81.80 | - | - | - |
| PAR [20] | 85.40 | 97.60 | 99.40 | **99.90** |
| Spindle [12] | 88.50 | 97.80 | 98.60 | 99.20 |
| PDC [21] | 88.70 | 98.61 | 99.24 | 99.67 |
| AACN [8] | **91.39** | **98.89** | 99.48 | 99.75 |
| Ours | 88.73 | 98.76 | **99.56** | 99.67 |

**Table 2**. Comparison results on Market1501.

| Market1501 | R-1 | mAP |
|---|---|---|
| SVDNet [19] | 82.30 | 62.10 |
| PAR [20] | 81.00 | 63.40 |
| Spindle [12] | 76.90 | - |
| PDC [21] | 84.14 | 63.41 |
| AACN [8] | 85.90 | 66.87 |
| Ours | **86.10** | **67.60** |

## 3. EXPERIMENTS

### 3.1. Datasets and Protocols

Our proposed RIN framework is evaluated on several public person ReID datasets namely Market1501 [1], CUHK03 [16] and DukeMTMC-reID [17]. In the standard evaluation protocol of Market1501 [1], the training set consists of 751 identities with a total of 12936 images while the test set consists of 750 identities containing 19734 gallery images and 3368 query images. In DukeMTMC-reID [17], the training set contains 702 identities with 16522 images while the test set consists of 702 identities containing 16522 gallery images and 2228 query images. The CUHK03 [16] consists of 13164 images with a total 1467 identities captured by 6 cameras. In the standard protocol, the training set contains 1160 identities while the test set contains 100 identities.

We evaluate the quality of our model using Cumulative Matching Characteristic (CMC) curves and mean average precision (mAP). All the experiments are performed in single query setting.

### 3.2. Implementation Details

The input images of our model is $256 \times 256$ pixels. PPGN network is trained on MPII [18] with over 25k images containing over 40k people with annotated body joints. Meanwhile a pretrained Resnet50 network is fine tuned on re-id datasets. Then we fix both network parameters, add other layers and train them. Finally, all modules are jointly fine-tuned.

### 3.3. Person re-identification performance

The proposed RIN is compared with recent approaches. These approaches are categorized into two sets: pose irrelevant and pose based methods. One set is the Singular Vector Decomposition method (SVDNet) [19], the Online Instance Matching method (OIM) [22], the pedestrian alignment network (PAN) [23], the Part-Aligned Representation (PAR) [20]. The other set utilized pose estimation network as a part, which includes the Spindle Net (Spindle) [12], the Pose-driven Deep Convolutional model (PDC) [21]. The experiment results are presented in Table [1, 2, 3]. It shows that our proposed RIN outperforms most approches on these datasets but the performance is lower than AACN in several datasets. It is mainly because the pose of these datasets changes little and the clothes in these datasets are mostly in simple color. Moreover, we measure the effectiveness of our proposed identical part area comparison method in Table 4 and give some masks generated by PPGN network in Fig.4.

**Table 3**. Comparison results on DukeMTMC-reID

| DukeMTMC-reID | R-1 | mAP |
|---|---|---|
| OIM [22] | 68.10 | - |
| PAN [23] | 71.59 | 51.51 |
| SVDNet [19] | 76.70 | 56.80 |
| AACN [8] | 76.84 | **59.25** |
| Ours | **77.20** | 56.9 |

**Table 4**. Effectiveness of identical area comparison

| Rank-1 | Market1501 | DukeMTMC |
|---|---|---|
| RIN-w/o-i | 85.1 | 73.87 |
| RIN-i | **86.10** | **77.20** |

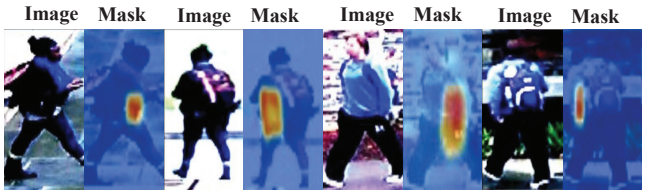


**Fig. 4**. Illustration of some identical part areas our model finds.

## 4. CONCLUSIONS

In this paper, we propose a Rotation Invariant Network (RIN) to deal with the self-occlusion and misalignment caused by human body horizontal rotation in person re-identification. RIN is composed of two main components, the Pose-guided Part Generation Network and Identical Area Comparison Mechanism, where PPGN is to generate part heatmap from coarse to fine and IACM is to compute the pairwise distance between cross-view intersection part areas. Extensive experiments demonstrate that our method achieves good performance on several public datasets.

## 5. REFERENCES


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

[2] Zheng Wang, Ruimin Hu, Chen Chen, Yi Yu, Junjun Jiang, Chao Liang, and Shin'ichi Satoh, "Person reidentification via discrepancy matrix and matrix metric," *IEEE Transactions on Cybernetics*, vol. 48, no. 10, pp. 3006–3020, 2018.

[3] Weijian Ruan, Jun Chen, Yi Wu, Jinqiao Wang, Chao Liang, Junjun Jiang, and Ruimin Hu, "Multi-correlation filters with triangle-structure constraints for object tracking," *IEEE Transactions on Multimedia*, 2018.

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

[5] Zheng Wang, Ruimin Hu, Yi Yu, Chao Liang, and Chen Chen, "Taichi distance for person re-identification," in *ICASSP*, 2017.

[6] Zheng Wang, Ruimin Hu, Chao Liang, Yi Yu, Junjun Jiang, Mang Ye, Jun Chen, and Qingming Leng, "Zero-shot person re-identification via cross-view consistency," *IEEE Transactions on Multimedia*, vol. 18, no. 2, pp. 260–272, 2016.

[7] Jin Wang, Zheng Wang, Chao Liang, Changxin Gao, and Nong Sang, "Equidistance constrained metric learning for person re-identification," *Pattern Recognition*, vol. 74, pp. 38–51, 2018.

[8] Jing Xu, Rui Zhao, Feng Zhu, Huaming Wang, and Wanli Ouyang, "Attention-aware compositional network for person re-identification," in *CVPR*, 2018.

[9] Alejandro Newell, Kaiyu Yang, and Jia Deng, "Stacked hourglass networks for human pose estimation," in *ECCV*, 2016.

[10] Eldar Insafutdinov, Leonid Pishchulin, Bjoern Andres, Mykhaylo Andriluka, and Bernt Schiele, "Deepercut: A deeper, stronger, and faster multi-person pose estimation model," in *ECCV*, 2016.

[11] Liang Zheng, Yujia Huang, Huchuan Lu, and Yi Yang, "Pose invariant embedding for deep person re-identification," *arXiv preprint arXiv:1701.07732*, 2017.

[12] Haiyu Zhao, Maoqing Tian, Shuyang Sun, Jing Shao, Junjie Yan, Shuai Yi, Xiaogang Wang, and Xiaoou Tang, "Spindle net: Person re-identification with human body region guided feature decomposition and fusion," in *ICCV*, 2017.

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

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

[15] Xuan Zhang, Hao Luo, Xing Fan, Weilai Xiang, Yixiao Sun, Qiqi Xiao, Wei Jiang, Chi Zhang, and Jian Sun, "Alignedreid: Surpassing human-level performance in person re-identification," *arXiv preprint arXiv:1711.08184*, 2017.

[16] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang, "Deepreid: Deep filter pairing neural network for person re-identification," in *CVPR*, 2014.

[17] Zhedong Zheng, Liang Zheng, and Yi Yang, "Unlabeled samples generated by gan improve the person re-identification baseline in vitro," *arXiv preprint arXiv:1701.07717*, 2017.

[18] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele, "2d human pose estimation: New benchmark and state of the art analysis," in *CVPR*, 2014.

[19] Yifan Sun, Liang Zheng, Weijian Deng, and Shengjin Wang, "Svdnet for pedestrian retrieval," in *ICCV*, 2017.

[20] Liming Zhao, Xi Li, Yueting Zhuang, and Jingdong Wang, "Deeply-learned part-aligned representations for person re-identification," in *ICCV*, 2017.

[21] Chi Su, Jianing Li, Shiliang Zhang, Junliang Xing, Wen Gao, and Qi Tian, "Pose-driven deep convolutional model for person re-identification," in *ICCV*, 2017.

[22] Tong Xiao, Shuang Li, Bochao Wang, Liang Lin, and Xiaogang Wang, "Joint detection and identification feature learning for person search," in *CVPR*, 2017.

[23] Zhedong Zheng, Liang Zheng, and Yi Yang, "Pedestrian alignment network for large-scale person re-identification," *IEEE Transactions on Circuits and Systems for Video Technology*, 2018.