Spindle Net: Person Re-identification with Human Body Region Guided Feature Decomposition and Fusion

Haiyu Zhao[♭♯], Maoqing Tian[♭], Shuyang Sun[♭], Jing Shao[♭], Junjie Yan[♭], Shuai Yi[♭],
Xiaogang Wang[♯], Xiaoou Tang[♯]

[♭]SenseTime Group Limited   [♯]The Chinese University of Hong Kong

{zhaohaiyu,tianmaoqing,sunshuyang,shaojing,yanjunjie,yishuai}@sensetime.com

{xgwang@ee.cuhk.edu.hk,xtang@ie.cuhk.edu.hk}

Abstract

Person re-identification (ReID) is an important task in video surveillance and has various applications. It is non-trivial due to complex background clutters, varying illumination conditions, and uncontrollable camera settings. Moreover, the person body misalignment caused by detectors or pose variations is sometimes too severe for feature matching across images. In this study, we propose a novel Convolutional Neural Network (CNN), called Spindle Net, based on human body region guided multi-stage feature decomposition and tree-structured competitive feature fusion. It is the first time human body structure information is considered in a CNN framework to facilitate feature learning. The proposed Spindle Net brings unique advantages: 1) it separately captures semantic features from different body regions thus the macro- and micro-body features can be well aligned across images, 2) the learned region features from different semantic regions are merged with a competitive scheme and discriminative features can be well preserved. State of the art performance can be achieved on multiple datasets by large margins. We further demonstrate the robustness and effectiveness of the proposed Spindle Net on our proposed dataset SenseReID without fine-tuning. [1]

1. Introduction

Person re-identification (ReID) aims at associating person images across cameras and temporal periods. Given one query image of one specific person, a person ReID system is expected to provide all the images of the same person from a large gallery database.

It is of great security interest and can be used for various surveillance applications. For example, when a child gets lost in a busy street, the ReID system can automatically

---

[1]H. Zhao and M. Tian share equal contribution. Source code and dataset can be found from https://github.com/yokattame/SpindleNet.

Figure 1. Challenges of person ReID. (a-b) Body region alignment across images. With human landmark information, the body regions shown in the blue boxes can be well aligned across images. However, without such information, directly matching based on location may results in ambiguities (the red boxes). (c) Detail information. With the local region features extracted from the blue boxes, detailed information can be captured and the two persons can be easily distinguished, even though their overall appearance are quite similar. (d) Occlusion. The lower-body region of the right image is occluded and the influence of the corresponding features should be weakened. All the examples are selected from the Market-1501 dataset [33].

search all the surveillance videos captured from nearby cameras and locate this child immediately. If an accident can be captured by one surveillance camera, all the witnesses of this accident can also be successfully tracked and located with a person ReID system, which can help the police find out more information of the accident and a lot of manpower can be saved.

Although the person ReID problem has been studied for years, it is still quite challenging. Firstly, human body regions cannot be well aligned across images. Two detection bounding boxes of the same person are shown in Fig. 1 (a), and the right box is not accurate. The head-shoulder region (blue box) of the left image is mis-aligned with the background region (red box) of the right image. The CNN

feature maps of these two regions should be quite different and cannot be directly compared. Even when the bounding boxes are accurate, the body region alignment is still quite challenging due to the large variations of human poses. An shown in Fig. 1 (b), the right leg region of the left image (blue box) is mis-aligned with the left leg region (red box) in the right image, where the black bag cannot be observed.

In most existing methods, features are extracted from the whole image [13, 29] or manually designed horizontal windows [14], thus the features cannot be well aligned and are not robust to detection errors or human pose variances. In our ReID pipeline, human landmark information is used, and body regions are obtained from a body region proposal network (RPN). Our ReID features can then be extracted from the seven body regions through a multi-stage ROI pooling pipeline. In this way, features of different body regions can be well aligned across images. As shown in Fig. 1, based on the body region information, the head-shoulder regions (blue boxes) in (a) and the right leg regions (blue boxes) in (b) can be well aligned, and their region features can be directly compared without ambiguity.

Secondly, some detail information is very important to discriminate different individuals. As shown in Fig. 1 (c), the general appearances of the two persons are quite similar. It is quite difficult to distinguish them if not focusing on the head region. However, most existing ReID methods adopt global features on the whole image [13, 29] and are very difficult to capture such detailed information. In our approach, with the region features, a lot of detail information can be better represented, which should be quite helpful for identifying individuals with slight differences.

Lastly, occlusion is another main challenge. Some body parts may be occluded in some images which makes the association process more difficult. How to separate the occluded features from the good ones is also very important for person ReID. As shown in Fig. 1 (d), the lower-body region of the right image is occluded by one white object, thus the feature importance of this region should be weakened in the feature comparison process.

However, no feature selection and filtering strategy is adopted in most existing ReID methods and all feature components are regarded equally useful. Actually, different body regions should have different importance when used for identifying persons. In order to make better use of the region features, a tree-structured feature fusion strategy is adopted in our approach instead of directly concatenating the region features together. Regions features of different semantic levels are merged separately in different stages. Micro features are merged in early stages and macro ones are merged later. Moreover, a competitive strategy is also used in the feature fusion process. With such strategy, good features can win the competition and thus can be preserved.

Compared with existing ReID approaches, our method

can achieve the state-of-the-art performance on most standard datasets with standard CMC evaluation metric. For example, our model can achieve an Top-1 accuracy of 76.9% on the Market-1501 [33] dataset, 11.0% higher than the the best result (65.9%) in literature [25]. In order to test the generalization ability of the learned model, we propose a new dataset (SenseReID) for testing purpose only. It is captured from real surveillance cameras and the person bounding boxes are obtained from state-of-the-art detection algorithm [20]. The proposed SenseReID dataset contains 1717 identities in total and our model can achieve much better performance compared with recent ReID approaches.

The contribution of this work can be summarized as the following three-folds. 1) It is the first time human body structure information is considered in a ReID pipeline. It can help align body region features across images and local detail information can be better described. 2) The Spindle Net is designed for the ReID task. Features of different body regions are first extracted by a multi-stage ROI pooling framework, and features of different semantic levels are pooled out separately at different stages. Then the regions features of different semantic levels are merged by a tree-structured fusion network with a competitive strategy. 3) A real surveillance ReID dataset, i.e. SenseReID, is proposed for performance evaluation purpose only. Our proposed method can achieve state of the art performance on the proposed dataset and multiple standard datasets.

2. Related Work

Most ReID pipelines are composed of two main steps, feature learning and metric learning. As for the feature learning step, traditional features [5, 6, 12, 14, 17, 21, 33] are widely used in existing pipelines. For example, an explicit polynomial kernel feature map is used to characterize the similarity information of all patch pairs between two images [6]. Liao *et al*. [14] proposed an effective feature representation called Local Maximal Occurrence (LOMO). With the great success of CNN features, a lot of recent ReID approaches [7, 13, 22, 24, 25, 28, 29] are designed based on CNN structure. For example, Li *et al*. [13] adopted a filter pairing neural network (FPNN) for person ReID. Cheng *et al*. [7] introduced a multi-channel CNN to learn body features from the input image. In [29], a pipeline is developed to learn generic feature representations from multiple domains and domain guided dropout is adopted to discard useless neurons for one specific domain. However, all the CNN features used in previous literature are extracted from the whole image or manually designed horizontal windows. Such features cannot be well aligned across images and cannot focus on each body part accurately. As introduced in Section 1, the accurate body region features should be very important to distinguish each individual.

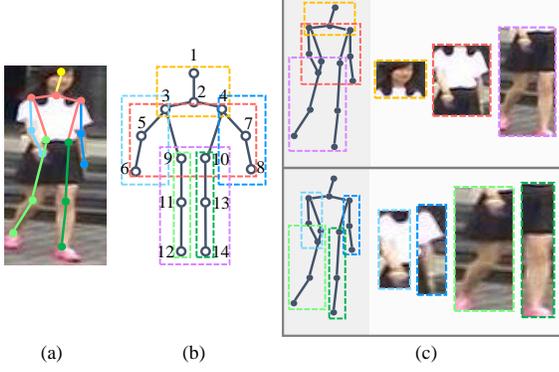As for the metric learning step, Liao *et al*. [14] pro-

Figure 2. Illustration of the Region Proposal Network. (a) One sample image and the fourteen body joints. (b) The fourteen body joints are assigned to seven sets. (c) The seven body sub-regions proposed by the RPN from the corresponding body joint sets.

posed a subspace based metric learning method called Cross-view Quadratic Discriminant Analysis (XQDA). In [11], Jose *et al.* introduced a metric learning formulation called Weighted Approximate Rank Component Analysis (WARCA). A logistic metric learning algorithm is developed by Liao *et al.* [15] with the PSD constraint and an asymmetric sample weighting strategy. Xiong *et al.* [31] proposed the four kernel-based distance learning approaches. Cheng *et al.* [7] proposed an improved triplet loss function that requires the intra-class feature distances to be less than the inter-class ones and a pre-defined threshold. In our paper, we mainly target on the feature learning step and Softmax classification loss is directly adopted for the metric learning step. Performance might be further improved if adopting advanced metric terms.

In recent years, the video based ReID problem is proposed and much work has been conducted on this field [4, 8, 18, 27]. In a video based ReID problem, sequential data is used instead of images. In this work, we only target on the image based ReID problem.

3. Body Region Proposal Network

As introduced in Fig. 1 (a-c), human landmark information can help align body regions across images and local details can be better described by the region features. Therefore, human body region information is included in the proposed ReID pipeline and the Region Proposal Network (RPN) is used to extract the body regions. Given an input image, the RPN generates seven rectangle region proposals representing seven sub-regions of the person body in the image, including the head-shoulder region, the upper body region, the lower body region, two arm regions, and two leg regions. The RPN contains two main steps, i.e. body joint localization and body region generation.

The first step of RPN is to locate the fourteen human body joints from one input image. A fully convolutional

architecture is adopted, which takes the human image as input and computes one response map $F_i \in \mathbb{R}^{X \times Y}$ ($i \in 1, ..., 14$) for each of the fourteen body joints. X and Y are the horizontal and vertical dimensions of the feature maps.

Inspired by the Convolutional Pose Machines (CPM) [26], we employ a sequential framework to generate these response maps in a coarse-to-fine manner. In each stage, a convolutional network is used to extract image features and then combine the response maps from the previous stage, to produce increasingly refined estimations for body joint locations. However, the model in CPM [26] is computational expensive. Three modifications are made to reduce model complexity yet produce high quality body joint response maps.

- Early convolution layers are shared among different stages in our framework to extract the image features.

- The pooling layers are replaced by convolution layers with stride 2 to reduce the feature map sizes.

- The input image size, the number of framework stages and the channel numbers of convolution layers are all reduced for fast computation.

The detailed structure and parameters of the RPN can be found in the supplementary material. With the feature maps, the fourteen body joints can be located by finding the coordinates with the maximum feature values.

$$P_i = [x_i, y_i] = \arg \max_{x \in [1, X], y \in [1, Y]} F_i(x, y) \qquad (1)$$

An example of the located fourteen body joints are visualized in Fig. 2 (a).

The second step of RPN is to obtain seven body sub-regions, including three macro sub-regions (head-shoulder, upper body, lower body) and four micro sub-regions (two arms, two legs) based on the body joint locations P_i.

As shown in Fig. 2 (b), the 14 located body joints are assigned to three macro sets $\mathcal{S}_1^A = [1, 2, 3, 4]$, $\mathcal{S}_2^A = [3, 4, 5, 6, 7, 8, 9, 10]$, $\mathcal{S}_3^A = [9, 10, 11, 12, 13, 14]$ and four micro sets $\mathcal{S}_1^B = [3, 5, 6]$, $\mathcal{S}_2^B = [4, 7, 8]$, $\mathcal{S}_3^B = [9, 11, 12]$, $\mathcal{S}_4^B = [10, 13, 14]$ according to the index numbers. For each of the body joint set $\mathcal{S} \in \{\mathcal{S}_1^A, \mathcal{S}_2^A, \mathcal{S}_3^A, \mathcal{S}_1^B, \mathcal{S}_2^B, \mathcal{S}_3^B, \mathcal{S}_4^B\}$, the corresponding sub-region bounding box $\mathcal{B} \in \{\mathcal{B}_1^A, \mathcal{B}_2^A, \mathcal{B}_3^A, \mathcal{B}_1^B, \mathcal{B}_2^B, \mathcal{B}_3^B, \mathcal{B}_4^B\}$ can be obtained based on the location coordinates of the all body joints in each macro/micro set.

$$\begin{aligned} \mathcal{B} &= [x_{\min}, x_{\max}, y_{\min}, y_{\max}] \\ &= [\min_{i \in \mathcal{S}}(x_i), \max_{i \in \mathcal{S}}(x_i), \min_{i \in \mathcal{S}}(y_i), \max_{i \in \mathcal{S}}(y_i)] \end{aligned} \qquad (2)$$

An example of the proposed seven body sub-regions are visualized in Fig. 2 (c).
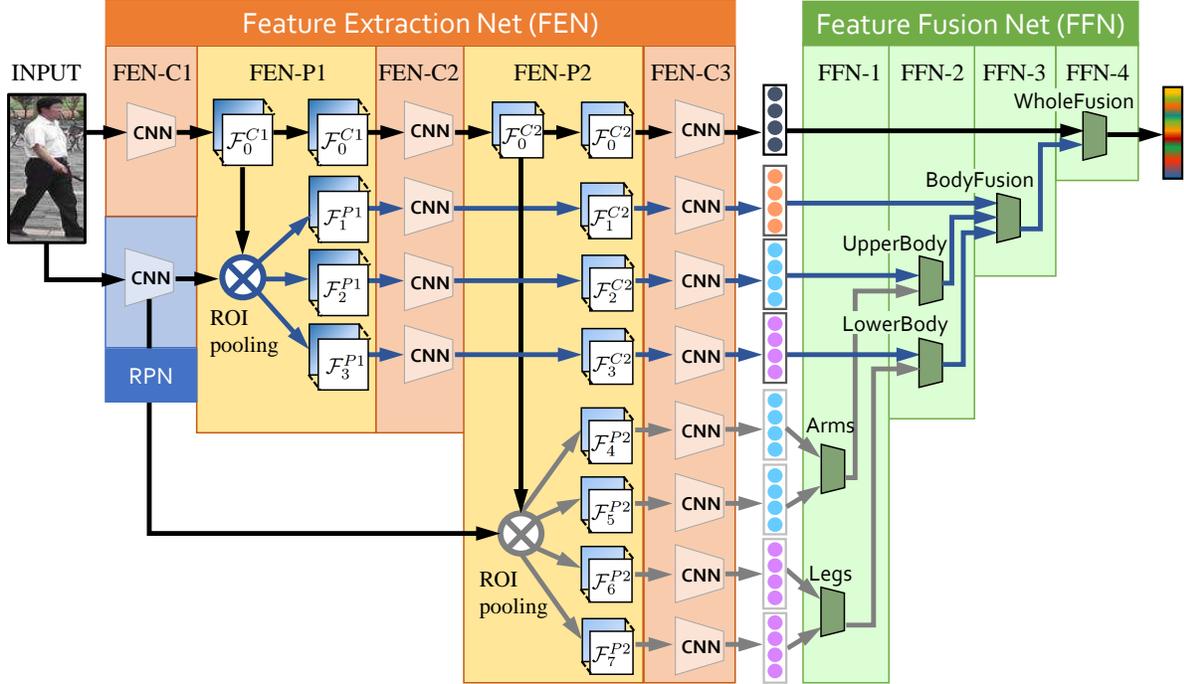
Figure 3. Flowchart of the proposed Spindle Net, including the Feature Extraction Net (FEN) and the Feature Fusion Net (FFN).

To train the RPN, the MPII human pose dataset [2] is used which covers various human poses. The target response map is generated as a Gaussian kernel around the ground truth body joint locations. The loss function at the output of each stage is defined as the L_2 distance between the estimated and the target response maps.

4. Body Region Guided Spindle Net

In this section, the proposed Spindle Net is introduced in detail which contains two main components, i.e. the Feature Extraction Network (FEN), and the Feature Fusion Network (FFN). The general flowchart of Spindle Net is shown in Fig. 3. The FEN takes the person image together with the region proposals as input and computes one global feature vector of the full image and seven sub-region feature vectors corresponding to the seven proposed body sub-regions. With the FFN, a final feature vector can be computed by merging the full image feature vector and the sub-region feature vectors together. The final feature vector can be used to distinguish different persons. More details of the two components are introduced in Sections 4.1 and 4.2. Training details are introduced in Section 4.3.

4.1. Feature Extraction Network (FEN)

With the FEN, one 256-dimensional feature vector can be extracted from each of the eight regions, including a full body region, and seven sub-regions proposed by the RPN, corresponding to three macro sub-regions and four micro

sub-regions. The FEN structure is shown in the left part of Fig. 3, which contains three convolution stages (FEN-C1, FEN-C2, FEN-C3) and two ROI pooling stages (FEN-P1, FEN-P2). The seven sub-regions are of different semantic levels. Therefore, the sub-region features are cropped from the full body feature maps at different stages, i.e. the three macro features are pooled out after the first convolution stage (FEN-C1) and the four micro features are pooled out after the second convolution stage (FEN-C2).

In FEN-C1, the input image is resized to 96×96 and convolved by three convolution layers and one inception module [23]. The output feature map of FEN-C1 is denoted as $\mathcal{F}_0^{C1}$ with spatial size 24×24. In FEN-P1, three feature maps $\mathcal{F}_i^{P1}(i = 1, 2, 3)$ are pooled out from $\mathcal{F}_0^{C1}$ based on the three macro sub-regions proposed by the RPN. The output size of FEN-P1 is 24×24. The three pooled feature maps $\mathcal{F}_i^{P1}(i = 1, 2, 3)$ are used as input of FEN-C2 together with $\mathcal{F}_0^{C1}$. In FEN-C2, four input feature maps pass through one inception module and output four output feature maps $\mathcal{F}_i^{C2}(i = 0, 1, 2, 3)$ of spatial size 12×12. These four output feature maps correspond to the full image region and three macro sub-regions. Afterwards, in FEN-P2, four feature maps $\mathcal{F}_i^{P2}(i = 4, 5, 6, 7)$ are pooled out from the full image feature map $\mathcal{F}_0^{C2}$ based on the four micro body sub-regions proposed by the RPN. The output size of FEN-P2 is 12×12. Together with $\mathcal{F}_i^{C2}(i = 0, 1, 2, 3)$, the four pooled feature maps $\mathcal{F}_i^{P2}(i = 4, 5, 6, 7)$ are also used as input of the last convolution stage, FEN-C3. In FEN-C3,
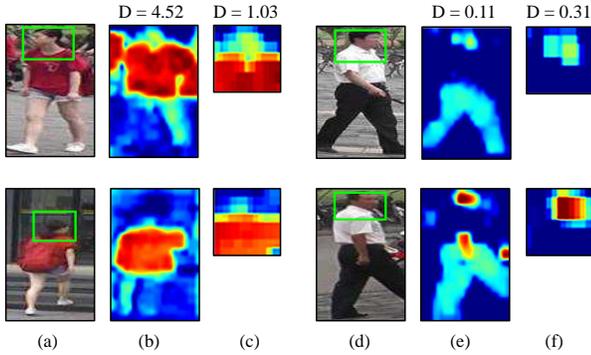
Figure 4. Two examples to demonstrate the effectiveness of the proposed sub-region features. (a) Two images of the same person. (b) Corresponding feature maps after FEN-C1. (c) Feature maps after FEN-P1. (D) Two similar persons. (b) Corresponding feature maps after FEN-C1. (c) Feature maps after FEN-P1. The average L_2 distances between feature maps are also listed.

each of the input feature maps first passes one inception module and the spacial dimensions are reduced to 6×6. Afterwards, one global pooling layer and one inner product layer are adopted to transform the inception result to a 256-dimensional feature vector. The output of FEN-C3 are eight 256-dimensional feature vectors, corresponding to the full image region, three macro body sub-regions, and four micro body sub-regions.

Two examples are shown in Fig. 4 to demonstrate the effectiveness of the proposed body sub-region features. In the first example, two images of the same person and the corresponding feature maps after FEN-C1 are shown in Fig. 4 (a-b). Due to the mis-alignment of the body regions, the feature maps in Fig. 4 (b) are quite different and the average L_2 distance of the two feature maps is $D = 4.52$. However, after FEN-P1, the pooled feature maps of the head-shoulder region (shown in Fig. 4 (c)) can be well aligned and the average distance is reduced to $D = 1.03$. It means that the pooled sub-region features can help better identify this individual. Another example is shown in Fig. 4 (d-f), the feature maps of two different persons shown in Fig. 4 (e) are quite similar. It is quite difficult to distinguish the two persons based on the feature maps of the full image. However, if we focus on the head-shoulder region (Fig. 4 (f)), the differences can be much more obvious.

4.2. Feature Fusion Network (FFN)

In the FFN, the eight feature vectors are combined together to generate one compact 256 dimensional feature vector that can well represent the whole image. A fusion unit is proposed for the feature fusion process, which takes two or more feature vectors of the same size as input and outputs one merged feature vector. The structure of FFN is shown in the right part of Fig. 3, and each fusion unit is represented by one green block.
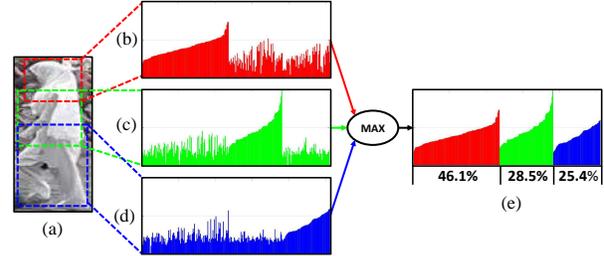

Figure 5. Illustration of feature fusion. Feature entries are sorted for better visualization. (a) Input image. (b-d) Three input feature vectors of the body fusion unit. The features of the head-shoulder region, the upper body region, and lower body region are marked in red, green and blue, respectively. (f) Result of the max operation. The head-shoulder features win 46.1% of the competition, much more than the other two region features in green and blue.

The fusion unit has two main processes. 1) The feature competition and selection process is conducted by an element-wise maximization operation. Feature entries with larger value can be preserved and smaller features are removed. 2) The feature transformation process is conducted by a inner product layer, so that the transformed result can be used for later fusion units.

Similar to the Feature Extraction Network where features of different body sub-regions are pooled out from different stages, the feature vectors of different body sub-regions are also merged in different stages based on their semantic levels and relationships. A tree-structured fusion strategy is proposed and features representing micro body sub-regions are merged in early stages and some macro features are merged in later stages.

As shown in Fig. 3, in the first stage, the features of the two leg regions, and the features of the two arm regions, are merged by two fusion units, separately. Afterwards, the two fusion results of the first stage are further merged with lower body features and upper body features, separately. Then, a fusion unit takes the two fusion results of the previous stage, together with the feature vector of the head-shoulder region as input and compute the merged feature vector of the whole body. Finally, the merged feature vector are concatenated with the feature vector of the full image, and are then transformed to the final 256-dimensional feature.

An example is shown in Fig. 5 to demonstrate the feature competition and fusion strategy based on the element-wise max operation. In this example, we focus on the fusion unit which takes three feature vectors, i.e. the head-shoulder features (red), the upper body features (green), and the lower body features (blue), as input and get the max responses. Even it is the same white color, wearing a white veil is much more distinguishable than wearing white clothes. Thus the head-shoulder features win much more competition in the max operation and 46.1% nodes are preserved in the output feature vector.

Figure 6. Eight datasets evaluated in our paper, including existing datasets, i.e. (a) CUHK03 [13], (b) CUHK01 [12], (c) PRID [10], (d) VIPeR [9], (e) 3DPeS [3], (f) i-LIDS [34], (g) Market-1501 [33], and (h) our proposed SenseReID testing dataset.

Dataset	#ID	#Trn/Val img	#Prb/Gal ID	#Prb/Gal img
CUHK03 [13]	1467	21012/5252	100/100	952/976
CUHK01 [12]	971	1552/388	485/485	972/972
PRID [10]	385	2997/749	100/649	100/649
VIPeR [9]	632	506/126	316/316	316/316
3DPeS [3]	193	420/104	96/96	246/316
i-LIDS [34]	119	194/48	60/60	98/130
Market-1501 [33]	1501	10348/2588	750/750	3368/15913
SenseReID	1717	0/0	522/1717	1040/3388

Table 1. The details of the eight datasets evaluated in our experiment. The number of train/validation images, together with the number of query/gallery identities and images are listed.

4.3. Training Details

A progressive strategy is adopted to train the Spindle Net, i.e. first training the Feature Extraction Net (FEN) and then the Feature Fusion Net (FFN). Model parameters are random initialized without pre-training.

The FEN is trained in the following three steps. 1) The full image branch is trained by giving supervision to the feature of the full image region. 2) Fixing the parameters in FEN-C1, the following three branches corresponding to the three macro sub-regions are trained by giving supervision to the three macro region features. 3) Fixing the parameters in FEN-C1 and FEN-C2, the last four branches corresponding to the four micro sub-regions are trained by giving supervision to the four micro region features. Then, the FFN are trained by giving supervision to the final feature vector.

When training FEN and FFN, existing ReID datasets are used and Softmax classification loss is adopted. The feature vector are transformed to a probability vector before computing the loss function.

5. Experiments

5.1. Datasets

To evaluate the performance of the proposed Spindle Net, seven existing standard person ReID datasets are evaluated in our experiments, including CUHK03 [13], CUHK01 [12], PRID [10], VIPeR [9], 3DPeS [3], i-LIDS [34], and Market-1501 [33]. In order to make fair comparison, we follow the standard data separation strategy as [33] for the Market-1501 dataset. For the other datasets, we follow the same settings as JSTL [29] to generate the training, validation, and test probe/gallery samples. The CUHK02 [12]

and PSDB [30] datasets are also used for training but not evaluated. In our implementation, the training and validation identities have no overlap with the test ones for all the datasets. Similar to JSTL [29], we merge all the training samples from these datasets to train a single ReID model which can be directly evaluated on all the testing datasets.

Besides existing datasets, the SenseReID dataset is proposed to evaluate the generalization ability of the trained ReID model. Unlike traditional datasets which are captured on campus or from some designed camera views, our proposed data are captured from twelve real world surveillance cameras. The proposed dataset contains no training data and is for testing purpose only. For the persons captured by more than one camera, the images of this individual are separated into the probe set and the gallery set according to the camera index. The probe set contains 522 identities, while the gallery set contains 1717 identities by including 1195 additional identities captured by only one camera. Selected samples of these datasets are shown in Fig. 6 and the statistical information of the datasets are listed in Table 1.

5.2. Comparison Results

The proposed Spindle Net is compared with several state-of-the-art ReID approaches in recent two years, including the metric learning method WARCA-χ^2 [11], the null space semi-supervised learning method NFST [32], the deep neural network PersonNet [28], the Siamese CNN with learnable gate S-CNN [25], the multi-domain CNN JSTL [29], the Joint Re-id method [1], the temporal adaptation model TMA [16], the hierarchical Gaussian model GOG+XQDA [17], the similarity learning method SCSP [5], the deep attribute model SSDAL+XQDA [22], the horizontal occurrence model LOMO-XQDA [14], the logistic metric learning model MLAPG [15], the triplet loss model TCP [7], the kernel-based distance learning approach MFA [31], the structured learning model CMC-top [19], and the bag of words model BoW-best [33]. The commonly used cumulative match curve (CMC) metric is adopted to quantitatively evaluate all the methods. One sample is random selected from duplicate gallery labels for all the evaluations.

CUHK03	Top-1	Top-5	Top-10	Top-20
WARCA-χ^2 [11]	**78.4**	94.6	-	-
NFST [32]	62.6	90.1	94.8	98.1
PersonNet [28]	64.8	89.4	94.9	98.2
S-CNN [25]	61.8	80.9	88.3	-
JSTL [29]	75.3	-	-	-
Spindle (Ours)	**88.5**	97.8	98.6	99.2

CUHK01	Top-1	Top-5	Top-10	Top-20
NFST [32]	69.1	86.9	91.8	95.4
PersonNet [28]	**71.1**	90.1	95.0	98.1
TCP [7]	53.7	84.3	91.0	96.3
JSTL [29]	66.6	-	-	-
Joint Re-id [1]	65.0	-	-	-
Spindle (Ours)	**79.9**	94.4	97.1	98.6

PRID	Top-1	Top-5	Top-10	Top-20
TMA [16]	54.2	73.8	83.1	90.2
NFST [32]	40.9	64.7	73.2	81.0
GOG+XQDA [17]	**68.4**	88.8	94.5	97.8
JSTL [29]	64.0	-	-	-
Spindle (Ours)	**67.0**	89.0	89.0	92.0

VIPeR	Top-1	Top-5	Top-10	Top-20
TMA [16]	48.2	-	87.7	95.5
NFST [32]	51.2	82.1	90.5	96.0
SCSP [5]	**53.5**	82.6	91.5	96.7
SSDAL+XQDA [22]	43.5	71.8	81.5	89.0
LOMO+XQDA [14]	40.0	-	80.5	91.1
MLAPG [15]	40.7	82.3	-	92.4
GOG+XQDA [17]	49.7	79.7	88.7	94.5
TCP [7]	47.8	74.7	84.8	91.1
Spindle (Ours)	**53.8**	74.1	83.2	92.1

3DPeS	Top-1	Top-5	Top-10	Top-20
WARCA-χ^2 [11]	51.9	75.6	-	-
SCSP [5]	**57.3**	79.0	-	91.5
MFA [31]	41.8	65.5	75.7	85.2
JSTL [29]	56.0	-	-	-
Spindle (Ours)	**62.1**	83.4	90.5	95.7

i-LIDS	Top-1	Top-5	Top-10	Top-20
WARCA-χ^2 [11]	36.6	66.1	-	-
CMC-top [19]	50.3	-	-	-
MFA [31]	32.1	58.8	72.2	85.9
TCP [7]	60.4	82.7	90.7	97.8
JSTL [29]	**64.6**	-	-	-
Spindle (Ours)	**66.3**	86.6	91.8	95.3

Market-1501	Top-1	Top-5	Top-10	Top-20
NFST [32]	55.4	-	-	-
PersonNet [28]	48.2	-	-	-
S-CNN [25]	**65.9**	-	-	-
BoW-best [33]	44.4	63.9	72.2	79.0
Spindle (Ours)	**76.9**	91.5	94.6	96.7

Table 2. Experimental results of the proposed Spindle Net and other comparisons on seven existing datasets. The CMC Top-1-5-10-20 accuracies are reported. The Top-1 accuracies of two best performing approaches are marked in bold. Due to space limitation, only some best performing methods are shown in the two tables.

SenseReID dataset	Top-1	Top-5	Top-10	Top-20
JSTL [29]	23.0	34.8	40.6	46.3
BoW-best [33]	22.4	-	-	-
Spindle (Ours)	**34.6**	52.7	59.9	66.7

Table 3. Experimental results of the proposed Spindle Net and other comparisons on our proposed SenseReID dataset. The CMC Top-1-5-10-20 accuracies are reported.

The testing procedure is repeated for 100 times and the average results are reported in Tables 2 and 3. From Table 2, we can observe that the proposed Spindle Net can achieve the best Top-1 accuracy on most existing ReID datasets. Especially, our model can achieve an 88.5 Top-1 accuracy on the CUHK03 dataset [13] , which is 10.1% higher than the best compared method [11]. Moreover, our model can achieve an Top-1 accuracy of 76.9% on the challenging Market-1501 dataset [33] , 11.0% better than the best comparison [25]. It is because our Spindle Net takes the body landmark information into consideration and region features are used in the ReID pipeline. Moreover, the feature selection and fusion strategy also helps obtain good compact features.

For the PRID dataset [10], our model achieves slightly lower Top-1 accuracy than GOG+XQDA [17]. It is also reasonable because there is only one trained model directly evaluated on all the datasets. The strong generalization ability may hurt the performance on one specific dataset. In order to further demonstrate the generalization ability, we test our model on the SenseReID dataset without fine-tuning and the results are listed in Table 3. We can observe that our model can get much better accuracy than the comparisons.

6. Investigations on Spindle Net

In this section, the proposed Spindle Net is thoroughly investigated, in terms of the Feature Extraction Network (FEN) and the Feature Fusion Network (FFN). The investigations are evaluated on the largest and the most challenging Market-1501 dataset [33].

6.1. Investigations on FEN

We first investigate the best position for pooling out the macro region features and the micro region features. There are three potential pooling positions, i.e. after FEN-C1, after FEN-C2, and after FEN-C3. For each position, the pooled macro and micro features are tested and the performance are reported in Fig. 7. We can observe that the feature maps generated by FEN-C1 are best for pooling out macro features while micro features should be pooled out from the FEN-C2 features. The macro regions contain more complex identity information, thus the macro features should be pooled out in earlier stage to have more independent learnable parameters.
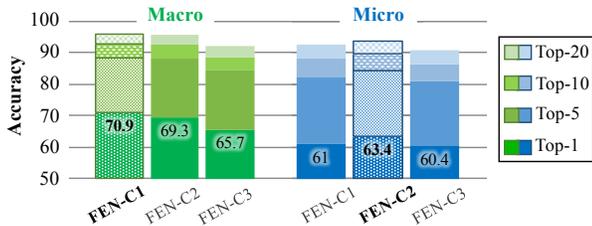
Figure 7. The evaluation performance of different pooling positions for the macro region features (left) and the micro region features (right) on Market-1501 [33] datasets.

	Market-1501	Top-1	Top-5	Top-10	Top-20
(a)	Full only	72.1	88.9	92.9	95.6
(b)	Full+FEN-C1/C1	74.3	90.5	94.1	96.5
	Full+FEN-C2/C2	73.1	90.0	93.8	96.2
	Full+FEN-C3/C3	67.8	85.9	90.6	93.9
(c)	**Full+FEN-C1/C2**	**74.7**	**90.8**	**94.3**	**96.6**
	Full+FEN-C1/C3	73.7	90.0	93.7	96.2
	Full+FEN-C2/C3	72.5	89.3	93.2	95.8
(d)	Full+FEN-C2/C1	74.0	90.5	94.1	96.5
	Full+FEN-C3/C1	72.2	89.4	93.3	95.9
	Full+FEN-C3/C2	72.0	89.2	93.2	95.9

Table 4. The evaluation performance of different combinations of the full image features and the macro/micro features extracted from different stages on Market-1501 [33] datasets. FEN-C1/C2 means the macro features are pooled from FEN-C1 and the micro features are pooled from FEN-C2.

Besides the effectiveness of the pooled features, the complementary property is also very important for selecting good feature combinations. Experiments are also conducted by joint testing different combination of the full image feature, the macro features, and the micro features. From the results in Table 4, we found that the combination of Full+FEN-C1/C2 achieves the best accuracies. Even without the Feature Fusion Network, the Top-1 accuracy can be improved by 2.6% to 74.7% by introducing the macro and micro region features compared with the results of Table 4 (a). Moreover, compared with the results in Table 4 (d), the better performance of (c) also demonstrates that macro features should be pooled out earlier than the micro features.

6.2. Investigations on FFN

There are two key factors of the proposed FFN, i.e., the tree fusion structure and the feature competition strategy. For the tree fusion structure, the results of using only one region feature are evaluated and listed in Table 5. From the results, we can observe that the full image feature can generally achieve better accuracies than the macro region features, while the macro region features are better than those micro ones. Thus the tree-structured fusion technology is adopted and better features are merged in later stages. On the other hand, such fusion structure is also consistent with

Market-1501	Top-1	Top-5	Top-10	Top-20
Full	72.1	88.9	92.9	95.6
Head and Shoulder	41.0	64.5	74.1	82.7
Upper Body	59.5	81.8	88.2	92.8
Lower Body	61.3	81.9	87.6	91.9
Right Arm	33.0	55.3	65.4	75.2
Left Arm	33.3	55.7	65.7	75.1
Right Leg	49.7	72.7	80.4	86.5
Left Leg	49.1	72.3	80.0	86.5

Table 5. The testing performance of each SINGLE feature component on the Market-1501 [33] datasets.

Market-1501	Top-1	Top-5	Top-10	Top-20
Linear + Concat.	72.8	89.1	93.0	95.6
Linear + Avg.	62.7	82.0	87.3	91.4
Linear + Max.	62.8	82.0	87.2	91.3
i-Tree + Concat.	66.5	86.4	91.3	94.7
i-Tree + Avg.	68.6	87.4	91.9	95.0
i-Tree + Max.	41.9	66.4	76.2	84.1
Tree + Concat.	67.1	84.7	88.9	92.1
Tree + Avg.	74.3	90.4	93.9	96.3
Tree + Max. (Ours)	**76.3**	**91.1**	**94.5**	**96.5**
Fine-tune (Ours)	**76.9**	**91.4**	**94.6**	**96.7**

Table 6. Comparison results of different fusion structures and competition strategies on Market-1501 [33] datasets.

the human body structure. The limb features are merged first while some higher level information are merged later.

The proposed fusion structure (Tree) is compared with some other possible ones, including the Linear structure (features are merged one by one) and the inverse tree (i-Tree) structure (macro features are merged first). We also compare the proposed competition strategy (Max.) with some baselines including direct concatenation (Concat.), and the element-wise average (Avg.). The performance of different fusion structures and competition strategies are reported in Table 6. The proposed FFN (Tree+Max.) achieves the best performance. Global fine-tuning the whole Spindle Net can further improve the performance.

7. Conclusion

In this paper, a novel Spindle Net is proposed for person ReID. Features of different body regions are separated by a multi-stage ROI pooling network and merged by a tree-structured fusion network. We show that the multi-level body features are informative to help align body regions across different images, and capture discriminative details of individuals. Strong capacity of the proposed feature competition and fusion network is also verified. State of the art performance can be achieved on multiple datasets and the robustness of the proposed Spindle Net is demonstrated on the proposed SenseReID dataset without fine-tuning.

References

[1] E. Ahmed, M. Jones, and T. K. Marks. An improved deep learning architecture for person re-identification. In *CVPR*, 2015. 6, 7

[2] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *CVPR*, 2014. 4

[3] D. Baltieri, R. Vezzani, and R. Cucchiara. 3dpes: 3d people dataset for surveillance and forensics. In *Proceedings of the 2011 joint ACM workshop on Human gesture and behavior understanding*, 2011. 6

[4] L. Bazzani, M. Cristani, A. Perina, M. Farenzena, and V. Murino. Multiple-shot person re-identification by hpe signature. In *ICPR*, 2010. 3

[5] D. Chen, Z. Yuan, B. Chen, and N. Zheng. Similarity learning with spatial constraints for person re-identification. In *CVPR*, 2016. 2, 6, 7

[6] D. Chen, Z. Yuan, G. Hua, N. Zheng, and J. Wang. Similarity learning on an explicit polynomial kernel feature map for person re-identification. In *CVPR*, 2015. 2

[7] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng. Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In *CVPR*, 2016. 2, 3, 6, 7

[8] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. Person re-identification by symmetry-driven accumulation of local features. In *CVPR*, 2010. 3

[9] D. Gray, S. Brennan, and H. Tao. Evaluating appearance models for recognition, reacquisition, and tracking. In *Proc. IEEE International Workshop on Performance Evaluation for Tracking and Surveillance (PETS)*, 2007. 6

[10] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof. Person re-identification by descriptive and discriminative classification. In *Scandinavian Conference on Image Analysis*, 2011. 6, 7

[11] C. Jose and F. Fleuret. Scalable metric learning via weighted approximate rank component analysis. *arXiv preprint arXiv:1603.00370*, 2016. 3, 6, 7

[12] W. Li and X. Wang. Locally aligned feature transforms across views. In *CVPR*, 2013. 2, 6

[13] W. Li, R. Zhao, T. Xiao, and X. Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, 2014. 2, 6, 7

[14] S. Liao, Y. Hu, X. Zhu, and S. Z. Li. Person re-identification by local maximal occurrence representation and metric learning. In *CVPR*, 2015. 2, 6, 7

[15] S. Liao and S. Z. Li. Efficient psd constrained asymmetric metric learning for person re-identification. In *ICCV*, 2015. 3, 6, 7

[16] N. Martinel, A. Das, C. Micheloni, and A. K. Roy-Chowdhury. Temporal model adaptation for person re-identification. In *ECCV*, 2016. 6, 7

[17] T. Matsukawa, T. Okabe, E. Suzuki, and Y. Sato. Hierarchical gaussian descriptor for person re-identification. In *CVPR*, 2016. 2, 6, 7

[18] N. McLaughlin, J. Martinez del Rincon, and P. Miller. Recurrent convolutional network for video-based person re-identification. 2016. 3

[19] S. Paisitkriangkrai, C. Shen, and A. van den Hengel. Learning to rank in person re-identification with metric ensembles. In *CVPR*, 2015. 6, 7

[20] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, 2015. 2

[21] Z. Shi, T. M. Hospedales, and T. Xiang. Transferring a semantic representation for person re-identification and search. In *CVPR*, 2015. 2

[22] C. Su, S. Zhang, J. Xing, W. Gao, and Q. Tian. Deep attributes driven multi-camera person re-identification. *arXiv preprint arXiv:1605.03259*, 2016. 2, 6, 7

[23] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *CVPR*, 2015. 4

[24] E. Ustinova, Y. Ganin, and V. Lempitsky. Multiregion bilinear convolutional neural networks for person re-identification. *arXiv preprint arXiv:1512.05300*, 2015. 2

[25] R. R. Varior, M. Haloi, and G. Wang. Gated siamese convolutional neural network architecture for human re-identification. In *ECCV*, 2016. 2, 6, 7

[26] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional pose machines. *arXiv preprint arXiv:1602.00134*, 2016. 3

[27] L. Wu, C. Shen, and A. v. d. Hengel. Deep recurrent convolutional networks for video-based person re-identification: An end-to-end approach. *arXiv preprint arXiv:1606.01609*, 2016. 3

[28] L. Wu, C. Shen, and A. v. d. Hengel. Personnet: Person re-identification with deep convolutional neural networks. *arXiv preprint arXiv:1601.07255*, 2016. 2, 6, 7

[29] T. Xiao, H. Li, W. Ouyang, and X. Wang. Learning deep feature representations with domain guided dropout for person re-identification. *arXiv preprint arXiv:1604.07528*, 2016. 2, 6, 7

[30] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang. End-to-end deep learning for person search. *arXiv preprint arXiv:1604.01850*, 2016. 6

[31] F. Xiong, M. Gou, O. Camps, and M. Sznaier. Person re-identification using kernel-based metric learning methods. In *ECCV*, 2014. 3, 6, 7

[32] L. Zhang, T. Xiang, and S. Gong. Learning a discriminative null space for person re-identification. *arXiv preprint arXiv:1603.02139*, 2016. 6, 7

[33] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person re-identification: A benchmark. In *ICCV*, 2015. 1, 2, 6, 7, 8

[34] W.-S. Zheng, S. Gong, and T. Xiang. Associating groups of people. In *BMVC*, 2009. 6