7th International Conference on Knowledge and Innovation in Engineering, Science and Technology
15 - 17 December, 2020 Berlin, Germany

Refined Granularity Extraction for Person Reidentification

Motaz Sabri
RidgeI, Tokyo, Chiyoda 161 Japan

Abstract

The usage of global and partial features has proven essential in person reidentification (Re-ID) tasks. Extracting both features does not happen uniformly resulting in representations that are either focusing on local representations or posture details. This consequently lowers the efficiency and robustness against scenarios with large variances. In this paper, we propose a feature learning strategy integrating discriminative information with refined granularities. We carefully design a multibranch deep network architecture with one expert branch for global feature representations and two expert branches for local feature representations. We replace focusing on semantic regions with a focus on several stripes of images partitions. The stripes number varies in different local branches to obtain local feature representations with multiple granularities. We evaluate our approach on four challenging datasets (Market1501, MSMT17, DukeMTMC Re-ID and CUHK03) where it achieves state of the art performance among both supervised and unsupervised methods.

Keywords: Discriminative features learning, Feature localization, Granularity learning, Metric learning, Person Reidentification.

1. Introduction

Person re-identification (Re-ID) is a formidable task considering it aims to retrieve a given person among all the gallery images captured across different cameras. The complexity of the problem comes from large variations of postures, occlusion cases, clothes, background clutter, misdetection, etc. Deep convolutional neural network (CNN) expanded the Re-ID performance with better discrimination and robustness in many scenarios. Recent deep Re-ID methods (Bai et al., 2018, Zhang et al., 2017, Sun et al., 2018, Qian et al., 2018) surpassed state of art identification rates and average precision matching human abilities in some complicated cases.

Recent researches proposed locating significant body parts from images to represent local information. A body part region contains a small percentage of local information from the whole body. It also contains distraction by other related or unrelated information outside the regions. Those distractions are filtered by locating salience regions allowing more focus on identities and complementing global features. Therefore, highlighting partial regions is key element in the learning process (Yao,2019, Li et al., 2017). This happens by either enhancing features by middle-level attention on salient parts (Zhao et al., 2017, Liu et al., 2017, Liu et al., 2019, Li et al., 2018) and feeding strong structural information such as empirical knowledge about human bodies (Sun et al., 2018, Zhang et al., 2017) or strong learning based pose estimation (Su et al., 2002, Zhao et al., 2017). What is common among these approaches is their sensitivity to pose or occlusion variations. They are also unable to generalize for all the discriminative parts as they focus on specific regions with fixed semantics. They consist of multiple components that are not end-to-end learning process, which complicates learning.

Using CNN, we can extract more fine-grained saliency for local features of smaller part regions. Those regions of high importance come from equally split stripe that do not necessary belong to regions with specific semantics. As the number of horizontal stripes increases, the granularity of discriminative responses becomes finer as shown in the work of (Wen et al., 2016). Inspired by this observation, we design the Refined Granularity Network (RGN), that consists of one global expert branch and two local expert branches with parameters from the 5th residual stage of the ResNet-50 (Hadsell et al., 2006) backbone. In each expert local branch of RGN, we split pooled feature maps into part regions varying the numbers of stripes to learn local feature presentations independently, motivated by the findings of (Wen et al., 2016, Wang et al., 2018). We used equally divided parts for local representation to design end-to-end learning process. Our experiments demonstrate the effectiveness of our training strategy on Market-1501 (Zheng et al., 2015), MSMT17 (Wei and Tian. et al., 2018), DukeMTMC-reid (Ristani et al., 2002) and CUHK03 (Li et al., 2014) datasets. Besides outperforming other methods, our approach also increases the robustness against viewpoint and scale changes due to its high feature localization ability.

The remainder of this paper is organized as follows. Related works are introduced in Section 2 and details of the proposed method are provided in Section 3. Experimental results are presented in Sections 4 and results discussion is in Section 5. Conclusion is given in Section 6.

2. Related Work

There are multiple instances of previous works that use CNN for embedding learning for person Re-ID tasks in the literature. (Li et al., 2014, Zhao et al., 2017) use Siamese network architecture to generate body part embedding to learn Re- ID. (Zheng et al., 2018) proposes ID-discriminative Embedding (IDE) with ResNet-50 backbones offering high performance levels that influenced many modern deep Re-ID systems for embedding extraction. In (Ahmed et al., 2015, Xiao et al., 2016), intermediate embedding of image pairs is computed to characterize links between local parts in a handcrafted designed mechanism. (Yi et al., 2014) attempts to enhance the generalization of Re-ID embedding extraction by introducing domain guided dropout. (Wang et al., 2018) brings re-ranking strategy into Re-ID obtaining better accuracy.

2.1 Feature localization

Latest embedding learning strategies have enhanced the Re-ID accuracy under variety of conditions. Many of these strategies have been designed to emphasize discriminative features and suppress irrelevant ones for better learning. Some works use the human part or pose information to infer the regions with high importance for extracting foreground embedding (Song et al., 2018, Kalayeh et al., 2018, Xu et al., 2018). Some works focus on learning based on spatial positions or channels in end-to-end frameworks (Liu et al., 2018, Zhao et al., 2017, Li et al., 2018, Wang et al., 2018, Yang et al., 2019). (Zhao et al., 2017) tackle the embedding extraction by a metric learning approach based on part-based alignment matching. (Wen et al., 2016) slices input images feature maps into several stripes with same vertical length.

(Bai et al., 2018) extends (Wen et al., 2016) work by merging slices of local embedding with LSTM network and combining with global embedding learned from classification and metric learning. In (Li et al., 2018), both spatial and channel attentions are adopted to modulate the embedding. In general, convolutional layers with limited receptive fields are used to learn the spatial information. In (Zhang et al., 2020a), they propose a relation-aware embedding extractor to globally learn attention by exploiting the pairwise relations and achieve improvement for person Re-ID in images. However, these systems achieve similar performances as human, giving less margin to surpass (Zhao et al., 2017).

In this paper, we propose to combine the local representations from parts of images to improve Re-ID effectiveness. In the literature of part-based learning, researchers utilize body structural information, region proposal methods and spatial attention to enhance the representation learning. In (Cheng et al., 2016, Li et al., 2017, Wen et al., 2016), local embedding representations are learned by slicing images into horizontal stripes according to verisimilitude dynamics of human body. (Sun et al., 2017, Zheng et al., 2015) used body structural information and pose estimation to enhance region cropping areas with semantics. (Felzenszwalb et al., 2008, Kingma, 2015) introduce region proposal methods that are used to locate semantic partitions without strongly learning-based predictors. Those attention-based methods have been utilized in some part-based approaches (Li et al., 2017, Zhang et al., 2018, Zheng et al., 2015) and further enhanced by (Li et al., 2018, Liu et al., 2019, Liu et al., 2017).

2.2 Joint learning strategy

In our proposed method, we use horizontal stripes as part regions for local embedding learning and outperform state of the art methods. We use joint learning strategy by using both softmax and triplet losses in our proposed method. The most common loss functions are classification losses and metric losses in deep Re-ID systems training. Softmax loss is almost the only choice of classification loss function that can be used individually (Ahmed et al., 2015, Li et al., 2017, Li et al., 2018, Liu et al., 2017, Wen et al., 2016, Yi et al., 2014, Zhang et al., 2018, Zheng et al., 2018) or paired with other losses (Bai et al., 2018, Cheng et al., 2016, Li et al., 2014, Zhao et al., 2017) for embedding extraction. Metric losses have many variants with different ranking metrics. Contrastive loss (Guo and Zhang. et al., 2017) focuses on maximizing the distances between inter-class pairs and minimizing it between intra-class pairs as used in Siamese networks (Xiao et al., 2016). Triplet loss (Hermans et al., 2017, Sarfraz et al., 2018) creates a margin between the inter and intra distances of triplets knowing one of them is an anchor sample. Many methods (Chen et al., 2017, Cheng et al., 2016, Hermans et al., 2017, Su et al., 2002) improve the learning and performance issues in metric learning via triplet loss.

3. Proposed method

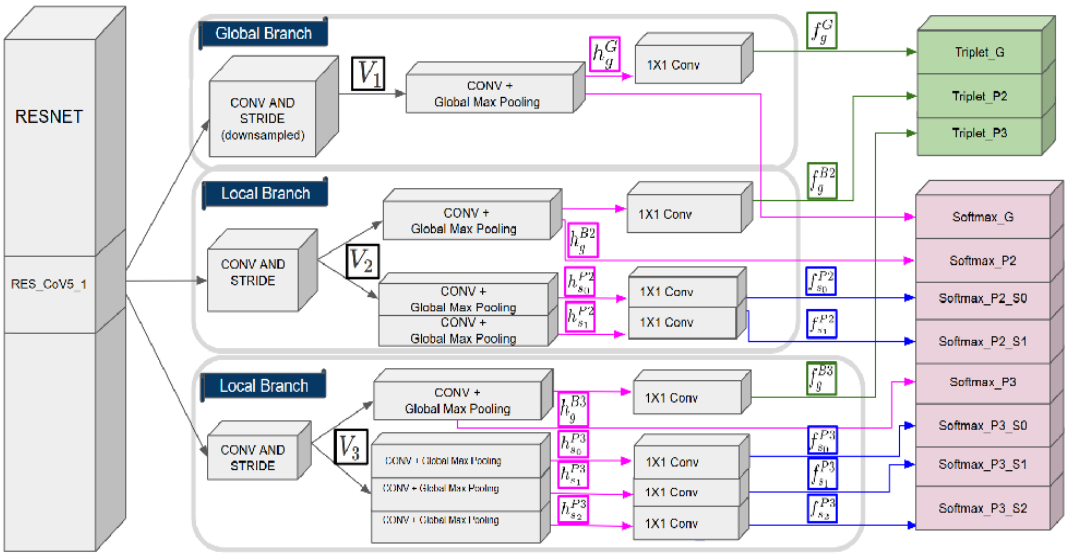
A deep CNN can gain discriminative abilities on different body parts based on their inherent semantic definition. Those abilities are enhanced by adding attention mechanism to highlight salient components (Zheng et al., 2018). To further marginalize the distraction of unrelated patterns in Re-ID targets in complicated scenes, higher activations must be concentrated on the main body of the subject. This is conventionally achieved in CNN by narrowing represented regions and learn local features by training via classification loss. As a result, responses on the local feature maps gradually cluster with variety of sizes on some salient semantic patterns.

The generation of the feature map requires analyzing high representative regions by focusing on specific patterns for representations and overcoming limitation of information in restricted regions (Wang et al., 2018). One major limitation in the feature extraction process is its unavoidable focus on basic granularity diversity regardless of empirical prior knowledge. As a result, it becomes intuitively hard to discriminate people identities based on the information brought by local regions. Using the classification learning strategy enforces the feature extraction to focus on more generalizable identity granularity, consequently pushing the training to explore useful refined details among relatively limited information. With enough level of granularity, CNN can capture details with most discriminative information.

3.1 Design

The architecture of RGN is shown in Figure 1. The feature encoder of our network is ResNet-50 which helps to achieve competitive performances in many ReID systems (Bai et al., 2018, Wen et al., 2016, Zhao et al., 2017). The features extracted after conv51 are fed as input to three independent branches. The upper branch of our network uses down-sampling with a stride 2 convolution layer in conv51 block, following a global max-pooling (GMP) (Almazan et al., 2018) on the corresponding output feature map and 1×1 convolution filter with ReLU activation and batch normalization (Szegedy, 2015) to reduce 2048 dim features h_g^G to 256 dim f_g^G . We refer to this branch as Global Branch as it learns features without partition information.

Figure 1: RGN architecture is shown with backbone getting forked from its intermediate residual blocks to one global and two local branches. During inference, all the reduced features, f_g^G, $f_g^{B2}, hf_g^{B3}, f_{s_i}^{P2}|_{i=1}^2, f_{s_i}^{P3}|_{i=1}^3$ are concatenated together as the final feature representation. Subbranches as well have no weight sharing. Every local feature has its own softmax loss function that represents an independent supervisory signal.



The other two branches share the same design with the global branch without down-sampling operations to maintain proper areas of reception fields for local features. We refer to the lower branches as local branches. The output feature maps of the local branches are split uniformly into several horizontal stripes. Branches two and three have the subbranch concept S_B which refers to the subbranch of the parent branch, subbranch contains partitions of the raw feature maps. During inference, the lower subbranches are concatenated and later combined with local and global features to form a comprehensive representation of the subject.

3.2 Loss function

Our choice of the loss function for the training phase is a combination of softmax loss for classification and triplet loss for metric learning. They are widely used in various deep Re- ID methods. We define our identification task as a multi-class classification problem. Therefore, the softmax of the i^{th} learned features f_i is formulated as:

$$\mathcal{L}_{Softmax} = \sum_{n=1}^{N} \log \frac{e^{W_{yn}^T f_i}}{\sum_{c=1}^{C} e^{W_c^T f_i}} \quad (1)$$

Where W_c denotes the weight vector for class c and N is the mini batch size while training for C classes. We follow the strategy of (Yao et al., 2019) in which bias terms are ignored in linear multi-class classifiers as this results in better discrimination performances. The softmax loss is used for the global features $h_g^G, h_g^{B2}, f_g^{B3}$ and local features after reduction $f_{s_i}^{P2}|_{i=1}^{2}, f_{s_i}^{P3}|_{i=1}^{3}$. After passing through the 1 ×1 convolution, the global features are reduced to form $f f_g^G . f_g^{B2} . f_g^{B3}$. and later trained with triplet loss to enhance ranking performances. Our triplet loss is the semi-hard triplet loss (Hermans et al., 2017). This triplet loss function $\mathcal{L}_T$ is formulated as:

$$\mathcal{L}_T = -\sum\sum \left\{ \mu + max_{1\le i\le Q} \left\| f_a^{(i)} - f_p^{(i)} \right\|_2 - min_{1\le i\le Q, 1\le j\le R, i\ne j} \left\| f_a^{(i)} - f_n^{(j)} \right\|_2 \right\} \quad (2)$$

Where $f_a^{(i)} f_p^{(i)} f_n^{(i)}$ are the features extracted from anchor, positive and negative samples receptively. Positive and negative refer to same and different person compared to anchor. The margin hyperparameter μ controls the differences of intra and inter distances. Considering we have R identities and K images from each identity, the triplet items are built by the furthest positive and closest negative sampled pairs. Inspired by the coarse-to-fine mechanism, we consider non-reduced features as coarse information and basis to learn classification and reduced features as fine information with learned metric. This is performed by feeding the reduced 256-dimension local features in second and third branches in addition to raw global pooled 2048-dimension from all branches to softmax losses while feeding triplet losses with the globally reduced features. This captures more salient details to meet the margin condition compared to softmax loss. As a result, when multiple granularities are fed to one mixed single branch, the features might dilute the importance of detailed information (Sun et al., 2015).

4. Experiments

In this section, we describe the configurations of our proposal, the data set we used for in our experiments, the evaluation methods we adopted and discussion of our results and findings.

4.1 Implementation details

We trained our model with input resized to 384 X 128 following (Sun et al., 2018). Our feature encoder of RGN is the ResNet- 50 pretrained on ImageNet (Deng et al., 2009). The global branch and local branches are all initialized with the same pretrained weights of the corresponding layers. To enhance network generalization, we used random horizontal flipping and cropping to images in the training dataset for data augmentation. We randomly selected Q image samples for R identities from the training set aligned with equation (2). Our experiments revealed that R = 8 and Q =4 yields good results. Triplet loss hyper parameter μ is set to 1.1 in all our experiments. We choose Adam (Kingma, 2015) as the network optimizer. We adopt the Adam optimizer to train all models for 600 epochs with the learning rate of 0.0008 and the weight decay of 0.0005. During inference time, we infer both the original image and its horizontal flip. Both embedding sets $f_g^G, f_g^{B2}, hf_g^{B3}, f_{s_i}^{P2}|_{i=1}^{2}, f_{s_i}^{P3}|_{i=1}^{3}$ are averaged to form the final features. All our experiments were executed on different datasets using the configuration above. We evaluate our proposed method using four Re-ID datasets: Market1501 (Zheng et al., 2015), MSMT17 (Wei and Tian. et al., 2018), DukeMTMC ReID (Zheng et al., 2017) and CUHK03 (Li et al., 2014).Using the DPM detector (Felzenszwalb et al., 2008), cropping bounding boxes are created for subjects. DukeMTMC ReID dataset is a subset of the DukeMTMC (Ristani et al., 2016) used for person Re-ID.

4.2 Evaluation

In our experiments we evaluated the cumulative matching characteristics (CMC) at rank1 and mean average precision (mAP) on previously mentioned datasets. On Market1501 and MSMT17 datasets following the common practices, our experiments included single and multiple query modes. On CUHK03 we followed the protocol used in (Zhong et al., 2017). We validate the effectiveness of our proposed method and show the comparisons in Table 1.

Table 1: Performance (%) comparisons with the state-of-the-art on CUHK03, Market1501 and MSMT17. Cumulative Matching Characteristics (CMC) is used as evaluation metrics for person reidentification.

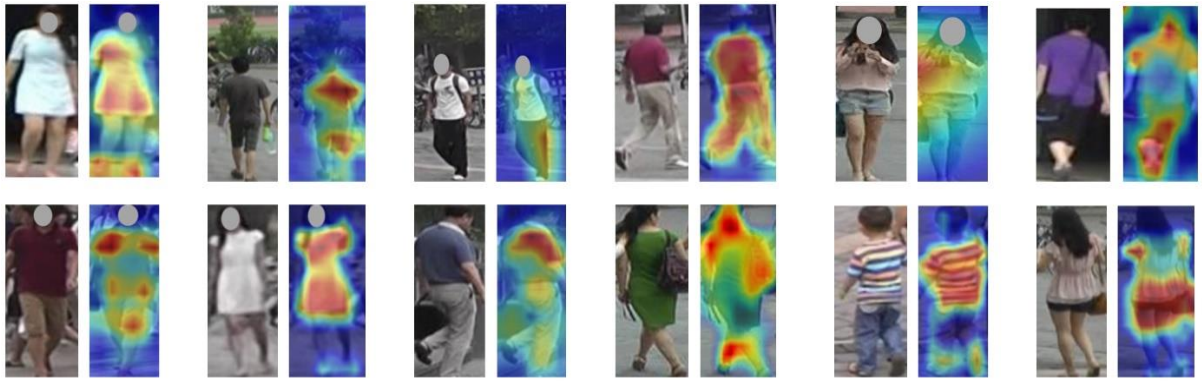
	CUHK03				Market 1501		MSMT17	
Model	Labeled		Detected					
	Rank1	MAP	Rank1	MAP	Rank1	MAP	Rank1	MAP
(Song 2018)	50.1	50.2	46.7	46.9	83.8	74.3	-	-
(Xu 2018)	-	-	-	-	85.9	66.9	-	-
(Kalayeh 2018)	-	-	-	-	92.5	81.3	-	-
(Li 2018)	44.4	41	41.7	38.6	91.2	75.7	-	-
(Si 2018)	-	-	-	-	91.4	76.6	-	-
(Wang 2018)	69	63.9	65.5	60.5	93.1	82.3	-	-
(Chen 2019)	77.2	72.4	71.7	65.4	95.1	85	-	-
(Fang 2019)	78.6	76.1	76.2	73.2	95.1	84.7	79.5	56.8
(Sun 2018)	63.7	57.5	-	-	93.8	81.6	68.2	40.4
(Fu 2019)	63.9	57.5	-	-	94.2	82.7	-	-
(Wang 2018)	68	67.4	66.8	66	95.7	86.9	-	-
(Hou 2020)	-	-	-	-	94.4	83.1	75.5	46.8
(Zheng 2019)	-	-	-	-	94.8	86	77.2	52.3
(Zheng 2019)	78.9	75.2	78.2	73.1	95.7	87.6	-	-
(Zhou 2019)	-	-	72.3	67.8	94.8	84.9	78.7	52.9
(Chen 2020)	86.8	**84.0**	**84.7**	81.0	95.7	88.5	**83.8**	58.5
(Zhang 2020)	81.1	77.4	79.6	74.5	96.1	88.4	80.3	57.5
RGN (Ours)	**86.9**	82.3	82.1	**81.1**	**96.8**	**90.3**	83.3	**60.1**

4.3 Ablation Study

In comparison with the approaches that leverage human semantics (Bai et al., 2018, Schroff et al., 2015, Wen et al., 2016) through attention on human part segmentation and those (Zheng et al., 2017, Sarfraz et al., 2018, Chen et al., 2017) which learn attention from input images themselves, our method significantly outperforms most of them. On CUHK03 (Detected), Market1501, and the largescale MSMT17 datasets, in comparison with all other approaches, RGN achieves the competitive performance which outperforms the second-best approach by 0.1%, 1.8% and 1.6% in mAP accuracy, respectively. According to Table 2, RGN method outperforms state of the art methods on the challenging DukeMTMCReID dataset and Market 1501 dataset. We apply the Grad-CAM (Selvaraju et al., 2017) for the qualitative analysis.

Figure 2 shows the generated masks of RGN covering the person regions accurately for some samples. Loss functions of our attention lead the network to focus on discriminative body parts in addition to global features. The attention focuses on the person and ignores the background. In comparison with the attention approach of (Woo et al., 2018) that does not consider relations between body parts, our attention focuses on regions with discriminative information.

Figure 2: GradCAM visualization according to gradient responses. Feature response maps extracted from output layers of every branches to form one attention map. Higher attention is under the redder regions.



	Rank1	mAP	Rank1	mAP
(Li et al., 2017)	85.1	65.5	89.7	74.5
(Zhang 2017	85.9	66.9	89.8	75.1
(Chen 2017)	88.6	72.6	92.2	80.4
(Chang 2018)	90	74.3	92.3	82.4
(Sohn. 2016)	90.1	75.3	-	-
(Sarfraz 2018)	90.4	80.5	-	-
(Li 2018)	91.2	75.7	93.8	82.8
(Si 2018)	91.4	76.6	-	-
(Schroff 2015)	92.7	82.5	-	-
(Chen 2017)	93.5	81.6	-	-
(Sun 2018)	93.8	81.6	-	-
(Wang 2018)	95.7	86.9	96.9	90.7
RGN(Ours)	**96.2**	**88.1**	**97.3**	**91.5**

(Sun 2017)	76.7	56.8
(Shen 2018)	80.7	66.4
(Si 2018)	81.8	64.6
(Sun 2018)	83.3	69.2
(Sarfraz 2018)	84.5	75.7
(Chen 2018)	84.9	69.5
(Almazan 2018)	85.2	72.8
(Chen 2020)	**91.0**	79.0
RGN (Ours)	89.9	**81.1**

Table 2: *Performance (%) comparisons with the state-of-the-art on Market 1501 at the left (with Single Query and Multiple) and DukeMTMCREID at the right.*

Figure 3 shows top-10 results for some given query pedestrian images from all datasets used in our experiments. The first row shows the great robustness regardless of the pose, brightness, or gait of given subject, RGN features can robustly represent discriminative information of their identities. The second query shows false positive case in which 2 people who look very similar are confused. The similarity is not only in clothing but also the posture. The third row holds an interesting observation in which the background has significant change in its content yet RGN

managed to react correctly to most of the results. The last row shows subject carrying a wearing coat being tracked correctly regardless of the posture or camera view.

Figure 3: Samples of query retrieval on all dataset using the proposed method. The first column contains the input images. The numbered images show the top10 retrieved images in descending order according to their similarity scores from left to right. The correct matches are marked with green rectangles.

5. Discussion

We performed additional ablation experiments with different settings on Market-1501 dataset in single query mode to evaluate RGN performance. To conduct fair comparison, we use the same set of configurations for RGN mentioned in experiment section and report the average performance with our settings.

5.1 RGN without triplet loss

In our experiment we utilize both the triplet loss and classification loss. Previous work show robustness in the joint training of those two loss functions (Bai et al., 2018, Cheng et al., 2016, Li et al., 2014, Zhang et al., 2017). When removing the metric loss from our proposal, the model acts like a standard ResNet-50 model. The effect reduces Rank-1 by -10.4% and mAP by -11.2%. We tried other ResNet versions such as ResNet-101 model. It can be observed that deeper feature extractor brings performance boost. Triplet loss helps to capture salient details to meet margin condition compared to softmax loss, matching results in (Schroff et al., 2015).

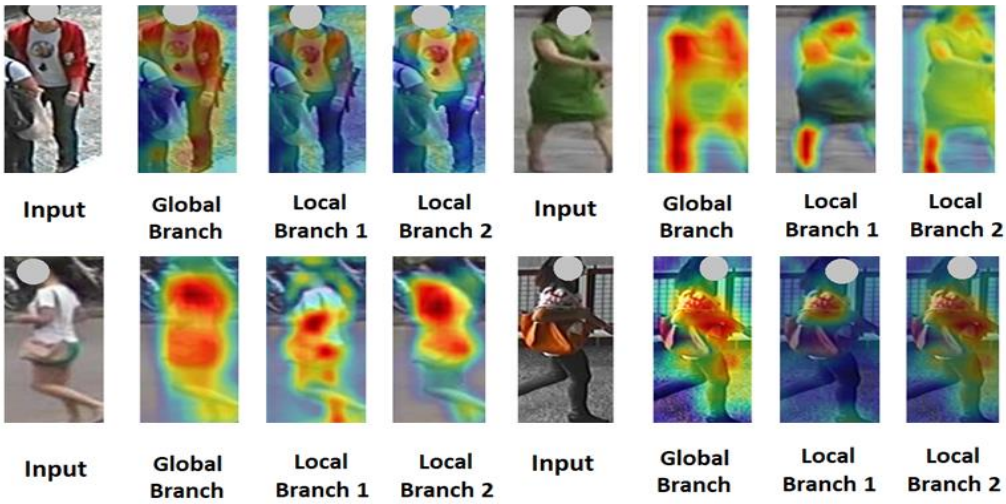
5.2 Feature learning

RGN model extracts both the global and local feature representations with multiple levels of granularities. Figure 4 shows the class activation maps for some input images, as seen from all the top of branches in RGN. Response maps filter out most of the background as it has no useful content for identification. The responses from global branch is mainly focused on main body parts while mostly excluding face due to the absence of identification information in such blur images. On local branches, the global responses on main body are ignored but additional attention is gained for more particular parts. The second branch shows more scattered responses on body parts highlighting pivotal semantic information.

6. Conclusion

We propose the Refined Granularity Network (RGN), multi-branch deep network for extracting features and learning discriminative representations in person re-identification tasks. Every branch becomes specialized in learning representation with certain granularity of body region. RGN learns local features on horizontally split feature stripes, which is completely end-to-end without any part locating operations. Beside outperforming most of state-of-the-art methods on several mainstream person Re-ID datasets, RGN generalizes well for many of the challenging cases in variety of datasets. We also performed ablation analysis providing more insight of the RGN model. Leveraging our unique exploitation of local and global features of human body images with inter and intra attention. It is in our interest to investigate more accurate and robust person ReID by incorporating trainable semantic human part parsing in our

Figure 4: GradCAM visualization according to gradient responses. Feature response maps extracted from output layers of every branches is shown. Themore red the region is the more attention is acquired in that region.



future work. It might also prove useful to investigate a learnable way to set the placement and size of the stripes proposed in this work.

References

A. Hermans, Beyer, and Leibe. (2017) In defense of the triplet loss for person reidentification. arXiv preprint arXiv:1703.07737

B. Chen, Deng, and Du. (2017) Noisy Softmax: Improving the Generalization Ability of DCNN via Postponing the Early Softmax Saturation. In CVPR pp. 5372-5381

C. Su, Li, Zhang, Xing, Gao, and Tian. (2002) Pose-driven Deep Convolutional Model for Person Reidentification. In ICCV pp. 3960-3969

C. Song, Huang, Ouyang, and Wang. (2018) Mask-guided contrastive attention model for person reidentification. In CVPR pp. 1179-1188

C. Wang, Zhang, Huang, Liu, and Wang. (2018) Mancs: A multitask attentional network with curriculum sampling for person reidentification. In ECCV pp. 384-400

D. Cheng, Gong, Zhou, Wang, and Zheng. (2016) Person Re-Identification by Multi-Channel Parts-Based CNN with Improved Triplet Loss Function. In CVPR pp. 1335-1344

D. Li, Chen, Zhang, and Huang. (2017) Learning Deep Context-aware Features over Body and Latent Parts for Person Reidentification. In CVPR pp. 384-393

D. Yi, Lei, Liao, and Li. (2014) Deep Metric Learning for Person Reidentification. In International Conference on Pattern Recognition pp. 34-39

D. Chen, Xu, Li, Sebe, and Wang. (2018) Group Consistent Similarity Learning via Deep CRF for Person ReIdentification. In CVPR pp. 8649-8658

E. Ahmed, Jones, and Marks. (2015) An improved deep learning architecture for person reidentification. In CVPR pp. 3908-3916

E. Hoffer and Ailon. (2015) Deep Metric Learning Using Triplet Network. In International Workshop on Similarity Based Pattern Recognition pp. 84-92

E. Ristani, Solera, Zou, Cucchiara, and Tomasi. (2002) Performance Measures and a Data Set for Multitarget, Multicamera Tracking. In ECCV pp. 17-35

F. Schroff, Kalenichenko, and Philbin. (2015) FaceNet: A Unified Embedding for Face Recognition and Clustering. In CVPR pp. 815-823

F. Yang, Yan, Lu, Jia, Xie, and Gao. (2019) Attention driven person reidentification. In CVPR pp. 143-155

G. Wang, Yang, Liu, Wang, Yang, Wang, Yu, Zhou and Sun. (2020) High Order Information Matters: Learning Relation and Topology for Occluded Person Re-Identification. In CVPR pp. 6449-6458

G. Wang, Y.Yuan, Chen, Li, Zhou. (2018) Learning Discriminative Features with Multiple Granularities for Person Re-Identification. In ICME pp. 274-282

H. Liu, Feng, Qi, Jiang, and Yan. (2019) Deep Residual Network with Self Attention Improves Person Re-Identification Accuracy. In ICMLC pp. 380-385

H. Yao, Zhang, Zhang, Li, and Tian. (2019) Deep Representation Learning With Part Loss for Person Re-Identification. In Transactions on Image Processing pp. 2860-2871

H. Zhao, Tian, Sun, Shao, Yan, Yi, Wang. (2017) Person Reidentification with Human Body Region Guided Feature Decomposition and Fusion. In CVPR pp. 907-915

H. Liu, Feng, Qi, Jiang, and Yan. (2018) End-to-end comparative attention networks for person reidentification. In Transactions on Image Processing pp. 3492-3506

J. Almazan, Gajic, Murray, and Larlus. (2018) ReID done right: towards good practices for person reidentification. arXiv preprint arXiv:1801.05339

J. Deng, Dong, Socher, L.J. Li, Li, and FeiFei. (2009) ImageNet: A largescale hierarchical image database. In CVPR pp. 248-255

J. Xu, Zhao, Zhu, Wang, and Ouyang. (2018) Attention-aware compositional network for person reidentification. In CVPR pp. 2119-2128

J. Si, Zhang, Li, Kuen, Kong, Kot. (2018) Dual attention matching network for context aware feature sequence-based person reidentification. In CVPR pp. 5363-5372

K. Kingma Ba, Adam: (2015) A Method for Stochastic Optimization. In ICLR pp. 3192-4001

J. Chen, Deng, and Hu. (2019) Mixed high order attention network for person reidentification. In ICCV pp. 371-381

K. Sohn. (2016) Improved Deep Metric Learning with Multiclass Npair Loss Objective. In Neural Information Processing Systems pp. 1857-1865

K. Zhou, Yang, Cavallaro, and T Xiang. (2019) Omni-scale feature learning for person reidentification. In ICCV pp. 3701-3711

L. Zhao, Li, Wang, and Zhuang. (2017) Deeply Learned Part Aligned Representations for Person Re-Identification. In ICCV pp. 3219-3228

L. Zheng, Shen, Tian, Wang, Wang, and Tian. (2015) Scalable person reidentification: A benchmark. In ICCV pp. 1116-1124

L. Wei, Zhang, Gao and Tian. (2018) Person Transfer GAN to Bridge Domain Gap for Person Reidentification. In CVPR pp. 7988 -8002

M. Kalayeh, Basaran, okmen, Kamasak, and Shah. (2018) Human semantic parsing for person reidentification. In CVPR pp. 1062-1071

M. Sarfraz, Schumann, Eberle, and Stiefelhagen. (2018) Pose-Sensitive Embedding for Person Re-Identification with Expanded Cross Neighborhood Re-Ranking. In CVPR pp. 420-429

P. Felzenszwalb, McAllester, and Ramanan. (2008) A discriminatively trained, multiscale, deformable part model. In CVPR 18

P. Fang, Zhou, Kumar Roy, Petersson, and Harandi. (2019) Bilinear attention networks for person retrieval. In ICCV pp. 8030-8039

R. Hadsell, Chopra, and LeCun. (2006) Dimensionality Reduction by Learning an Invariant Mapping. In CVPR pp. 1735-1742

R. Hou, Ma, Chang, Gu, Shan, and X Chen. (2020) Interaction and aggregation network for person reidentification. In CVPR pp. 9317-9326

R. Selvaraju, Cogswell, Das, Parikh, and Batra. (2017) Gradcam: Visual explanations from deep networks via gradient based localization. In ICCV pp. 618-626

S. Ioffe and Szegedy. (2015) Batch normalization: accelerating deep network training by reducing internal covariate shift. In ICML pp. 448-456

S. Woo, Park, Lee, and Kweon. (2018) Cbam: Convolutional block attention module. In ECCV pp. 319 -325

T. Xiao, Li, Ouyang, and Wang. (2016) Learning Deep Feature Representations with Domain Guided Dropout for Person Reidentification. In CVPR pp. 1249-1258

W. Chen, Chen, Zhang, and Huang. (2017) Beyond Triplet Loss: A Deep Quadruplet Network for Person Reidentification. In CVPR pp. 1320-1329

W. Li, Zhao, Xiao, and Wang. (2014) DeepReID: Deep Filter Pairing Neural Network for Person Reidentification. In CVPR pp. 152-159

W. Li, Zhu, and Gong. (2017) Person reidentification by deep joint learning of multiloss classification. In International Joint Conference on Artificial Intelligence pp. 2194-2200

W. Li, Zhu, and Gong. (2018) Harmonious Attention Network for Person Re-Identification. In CVPR pp. 2285-2294

X. Bai, Yang, Huang, Dou, Yu, and Xu. (2018) DeepPerson: Learning Discriminative Deep Features for Person Re-Identification. arXiv:1711.10658

X. Chang, Hospedales, and Xiang. (2018) Multilevel Factorisation Net for Person Reidentification. In CVPR pp. 2109-2118

X. Liu, Zhao, Tian, Sheng, Shao, Yi, Yan, and Wang. (2017) HydraPlusNet: Attentive Deep Features for Pedestrian Analysis. In ICCV 19

X. Qian, Fu, Wang, Xiang, Wu, Y.G. Jiang, and Xue. (2018) Pose-Normalized Image Generation for Person Reidentification. In ECCV pp. 661-678

X. Zhang, Luo, Fan, Xiang, Sun, Xiao, Jiang. (2017) Aligned ReID: Surpassing Human Level Performance in Person Re-Identification. arXiv:1711.08184

X. Chen, Fu, Zhao, Zheng, Song, Ji, Yang. (2020) Salience Guided Cascaded Suppression Network for Person Reidentification. In CVPR pp. 3300-3310

Y. Chen, Zhu, and Gong. (2017) Person Reidentification by Deep Learning Multiscale Representations. In Conference on Computer Vision Workshops pp. 2590-2600

Y. Zhang, Xiang, T.Hospedales, Lu. (2018) Deep Mutual Learning. In CVPR pp. 4320-4328

Y. Sun, Zheng, Deng, and Wang. (2017) SVDNet for Pedestrian Retrieval. In ICCV pp. 3820-3828

Y. Sun, Zheng, Yang, Tian, and Wang. (2018) Beyond Part Models: Person Retrieval with Refined Part Pooling. In ECCV pp. 234-778

Y. Wen, Zhang, Li, and Qiao. (2016) A Discriminative Feature Learning Approach for Deep Face Recognition. In ECCV pp. 499-515

Y. Fu, Wei, Zhou, Shi, Huang, Wang, Yao, and Huang. (2019) Horizontal pyramid matching for person reidentification. In AAAI pp. 8295-8302

Y. Shen, Li, Xiao, Yi, Chen, and Wang. (2018) Deep Group Shuffling Random Walk for Person Re-Identification. In CVPR pp. 2265-2274

Y. Sun, Wang, and Tang. (2015) Deeply learned face representations are spars. In CVPR pp. 2892-2900

Z. Zheng, Zheng, and Yang. (2018) Pedestrian alignment network for largescale person reidentification. In TCSVT pp. 3037-3045

Z. Zheng, Zheng, and Yang. (2017) Unlabeled Samples Generated by GAN Improve the Person Reidentification Baseline in Vitro. In ICCV pp. 3774-3782

Z. Zhong, Zheng, Cao, and Li. (2017) Reranking Person Reidentification with k Reciprocal Encoding. In CVPR pp. 3652-3661

Z. Zheng, Yang, Yu, Zheng, Yang, and Kautz. (2019) Joint discriminative and generative learning for person reidentification. In ICCV pp. 2138-2147

Z. Zheng, Yang, Yu, Zheng, Yang, and Kautz. (2019) Densely semantically aligned person reidentification. In ICCV pp. 667-676

Z. Zhang, Lan, Zeng, Jin, Chen. (2020) Relation Aware Global Attention for Person Reidentification. In CVPR pp. 3186-3195