Learning Discriminative and Generalizable Representations by Spatial-Channel Partition for Person Re-Identification

Hao Chen, Benoit Lagadec, Francois F Bremond

Learning Discriminative and Generalizable Representations by Spatial-Channel Partition for Person Re-Identification

Hao Chen[1,2], Benoit Lagadec[2], and Francois Bremond[1]
[1]University of Côte d'Azur, Inria, Stars Project-Team, France
{hao.chen,francois.bremond}@inria.fr
[2]European Systems Integration, France
benoit.lagadec@esifrance.net

Abstract

In Person Re-Identification (Re-ID) task, combining local and global features is a common strategy to overcome missing key parts and misalignment on models based only on global features. Using this combination, neural networks yield impressive performance in Re-ID task. Previous part-based models mainly focus on spatial partition strategies. Recently, operations on channel information, such as Group Normalization and Channel Attention, have brought significant progress to various visual tasks. However, channel partition has not drawn much attention in Person Re-ID. In this paper, we conduct a study to exploit the potential of channel partition in Re-ID task. Based on this study, we propose an end-to-end Spatial and Channel partition Representation network (SCR) in order to better exploit both spatial and channel information. Experiments conducted on three mainstream image-based evaluation protocols including Market-1501, DukeMTMC-ReID and CUHK03 and one video-based evaluation protocol MARS validate the performance of our model, which outperforms previous state-of-the-art in both single and cross domain Re-ID tasks.

1. Introduction

Person re-identification (Re-ID) targets at searching people across non-overlapping surveillance cameras by matching person images captured by different cameras. There are still various challenging problems to be solved in a real-world Re-ID task, such as camera view point changes, illumination differences, pose variation and partial occlusion. With the rapid development of deep learning based techniques, recent research with convolutional neural networks [18, 40] get remarkable advances in person Re-ID, which surpass the performance of traditional handcrafted methods [19, 39] on large datasets.

To measure similarities between two captured images, we need to build an appearance representation for each sample in datasets. The most intuitive method for building representations is to extract directly global feature map from the entire bounding box. However, Re-ID methods that rely solely on global features of a person are prone to errors in case of occlusion and misalignment. On the other hand, key local features (carried objects and body parts, such as face and hands) can not be always observable due to low camera resolution and occlusions.

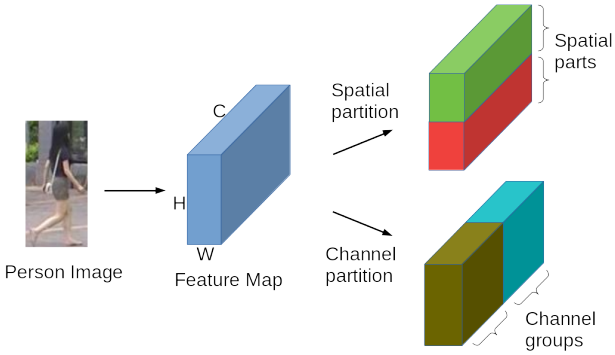

Figure 1. Example of a spatial-channel partition. H, W and C stand for respectively Height, Width and Channel in a deep feature map. In this example, we partition a whole feature map into two spatial parts (upper body and lower body) and two channel groups.

Since viewpoint change, partial occlusion and misalignment are frequent in real-world Re-ID task, complementing global features with local features addresses these issues and builds better person representations. This further improves a neural network's capacity to distinguish similar people based on small part differences.

Therefore, part-based models [1, 8, 22] have attracted a lot of attention in Person Re-ID research community. Recently, several sophisticated models [7, 30] have combined multiple partitions conducted along the height dimension in a pyramidal structure. They have significantly outperformed previous state-of-the-art. A feature map extracted from an image has 3 dimensions, *i.e.*, height, width and channel (also called depth). Because height and width dimensions correspond both to the spatial coordinates of pix-

els in an image, the partition conducted along the height and width dimensions is called spatial partition. Independently from spatial coordinates, partition conducted along the channel dimension is called channel partition. An example is shown in Figure 1.

Spatial partition is a common strategy in Re-ID task, which enables part-to-part matching by extracting local features corresponding to specific body parts. Channel information comes from filters in a convolutional layer. As the CNN goes deeper, last layers are more abstract, and outputs of the channels are higher level features, corresponding potentially to concepts such as hair color, body shape, *etc*. Channel partition does not extract local features from specific body parts, but keeps features that indicate the presence of these high level concepts. By splitting channels into several groups and training separate channel groups, semantic concepts in each channel group can be decorrelated. Therefore, channel partition allows to conduct semantic concept-to-concept matching (*e.g*., with or without a bag strap), which can complement spatially partitioned representations. As proven in [31, 20], semantic attributes show a strong generalibility in cross domain Re-ID task. By combing part-to-part and semantic concept-to-concept matching, we are able to build discriminative and generalizable representations with spatial-channel partition.

Another well-considered strategy for extracting discriminative features is attention mechanism. Attention mechanism helps CNNs to focus on the most discriminative part (called primary information) in feature maps. Features on other parts which are salient but less discriminative (called secondary information) are then neglected. These features can be complementary clues for distinguishing people with similar appearances. Partitions enable a CNN not only to consider primary features but also to keep secondary features.

In this work, we focus on how to build robust person representations for Re-ID task by complementing global features with local features extracted through partitions.

In summary, our contribution is twofold:

1. We conduct a comparative study between spatial, channel partitions and attention mechanism. Results of this study can be summarized by 2 statements: (a) Attention mechanism may remove useful secondary information, which can be kept by partitions. (b) Compared to traditional spatial partition, channel partition shows a superior capacity of maintaining secondary local information.

2. Spatial and channel partitions are combined (called spatial-channel partition) to further enhance deep neural networks' ability to learn secondary information. By adopting multiple spatial-channel partitions in a pyramidal structure, we propose a unified end-to-end trainable framework for Person Re-ID.

Our proposed framework is exhaustively evaluated on three image-based Re-ID datasets, Market-1501, DukeMTMC-ReID, CUHK03 and one video-based dataset MARS. On the MARS dataset, partition is also applied on the temporal dimension to build a more robust representation for each tracklet. The evaluation results show that our method can build both discriminative and generalizable representations, which outperform previous state-of-the-art in both supervised single domain and unsupervised cross-domain Re-ID tasks.

2. Study of Appearance Representations

Building discriminative appearance representations to measure quantitatively the similarity between query and gallery images is a common approach in Re-ID task. First, we evaluate robustness of appearance representations within the state-of-the-art. Then, we explain why spatial and channel partitions should be combined together and why they can outperform attentive models.

2.1. State-of-the-art

The two main approaches to make appearance representations robust in the state-of-the-art consist of choosing appropriate loss function and partitioning a person image into several spatial parts.

Loss Functions for Re-ID. A domain guided dropout is proposed in [33] to train a classification model on various datasets, which consider Re-ID as an identification task. Zheng *et al*. [40] use both pair-wise verification and classification loss to learn a more discriminative representation for Re-ID task. However, classification loss fails in cases of people wearing similar clothes, misaligned bounding boxes, *etc*. In [10], Batch Hard triplet loss is proposed to focus more on these hard samples. But the performance of triplet loss highly relies on how to select the hardest positive and negative pairs in a batch, which is difficult based on local features extracted from body parts. More details about sampling and hard pair selection are given in section 3.2. To get a better performance, we adopt triplet and classification losses to train our SCR model in a joint learning manner.

Spatial Part Based Models. Partitioning the entire body image into several spatial parts has always been a popular strategy in Re-ID task. Gray *et al*. [8] first propose to partition the person image into six equally-sized horizontal stripes and extract color and texture features in each stripe. Farenzena *et al*. [5] segment entire images into three salient and meaningful regions (head, torso and legs) by exploiting asymmetry and symmetry principles. These hand-crafted feature based approaches work well on small datasets, but are less robust and fail on large datasets.

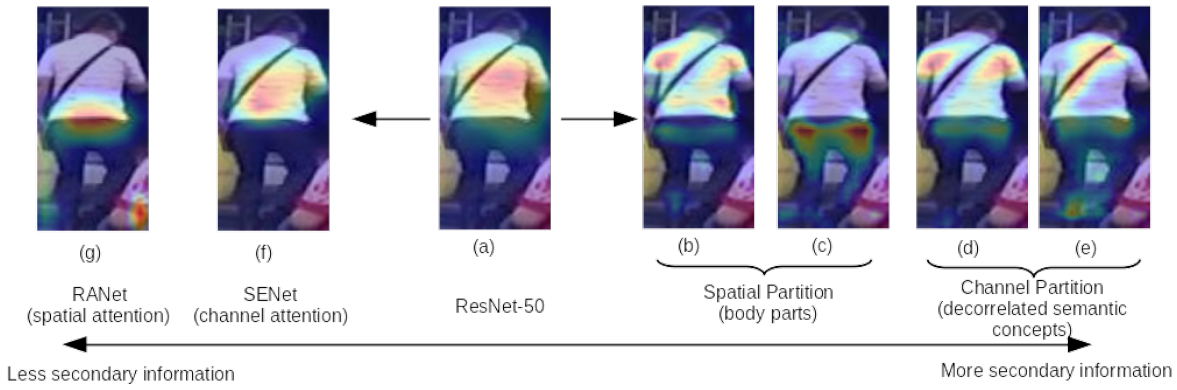

Figure 2. Comparisons of saliency maps generated by Grad-CAM [25] applied on 4 CNN models on Market-1501 test set. (a): A ResNet-50 w/o partitions nor attention mechanism. (b) to (e): A ResNet-50 w/ spatial-channel partition, where (b) and (c) are saliency maps on two spatial parts after spatial partition, (d) and (f) are saliency maps on two channel groups after channel partition. (f): Squeeze-and-Excitation Network [12]. (g): Residual Attention Network [29].

Recently, part-based deep learning methods build more robust representations with deep features learned on large datasets. Yao *et al*. [36] train CNN in several maximally activated regions on feature maps. In [4], authors propose a spatial-channel loss to ensure that each channel in the representation pays attention to a dedicated partitioned part of the body. But original channel information in the feature map is replaced by spatial information, making this loss inappropriate to maintain specifically channel information. Part-based Convolutional Baseline (PCB) proposed in [27] introduces a simple yet effective model based on six identical horizontal stripes. But same body part can be found in different stripes between different training samples, especially when bounding boxes are misaligned. To address this issue, a recent research trend is to combine multiple partitions to build a robust appearance representation.

Multi-partition Pyramidal Models. In HPM [7], Fu *et al*. partition respectively the entire body feature map into one, two, four and eight identical horizontal stripes. In MGN [30], authors split last layers of a ResNet into three branches and partition feature maps into one, two and three horizontal stripes. CPM [37] adopts multiple overlapping partitions. However, experiments show that these overlapping partitions do not increase the performance of our network. All these methods only consider multiple spatial partitions and neglect the channel ones.

Channel Group Operations. In AlexNet [14], Krizhevsky *et al*. firstly partition channels into 2 groups and introduce grouped convolutions to distribute a model into two GPUs. In ResNeXt [34], authors show that a larger number of channel groups can improve accuracy in image classification task without increasing computational complexity. MobileNet [11] adopts channel-wise convolutions where the number of groups equals the number of channels. In a similar way, by partitioning channels into several groups and computing within each channel group the mean and the variance for normalization, Group Normalization [32] outperforms Batch Normalization [13]. These studies confirm the potential of partitioned channel groups as an effective dimension in various visual tasks. In the following subsection, we discuss using partitioned channel groups to enhance the robustness of representations in Re-ID task.

2.2. Study of Partitioned Representations

In Figure 2 (a), the heat map generated by Grad-CAM [25] shows that a vanilla ResNet-50 trained with cross-entropy classification loss focuses on upper body especially on the region next to right arm, where other regions, *e.g*., right arm, legs and shoes are totally neglected. When training in this way, a CNN solely considers features on some discriminative regions. In consequence, it suffers from over-fitting on these regions and becomes less robust for hard samples. To overcome this issue and build a more generalized appearance representation, we employ spatial-channel partitions and train multiple classifiers separately on partitioned maps. Specific local features are fed separately into dedicated classifiers, each of them can be regarded as a local expert. A local expert works better on a dedicated part. Combining all local experts allows to build more robust representations for Re-ID. To verify this idea, we have conducted experiments on Market-1501 dataset, whose results are shown in Figure 2 (b) to (e). With spatial partition, more regions in upper and lower body are highlighted respectively in Figure 2 (b) and (c) as compared to (a). With channel partition, the model does not train local experts on dedicated body parts but on a group of high level features. Thanks to channel partition, in Figure 2 (d) and (e), the obtained saliency maps have more highlighted regions corresponding to semantic concepts, such as shoulder strap and shoes. The T-shirt is highlighted in (d), while the

shoulder strap and shoes are highlighted in (e). Different activated semantic concepts in (d) and (e) show that channel partition is able to decorrelate high level features in different channel groups and conduct finer concept-to-cencept matching.

Since activated regions of channel partitions are different from those of spatial partitions, we can infer that local features extracted from both types of partitions are complementary when trained jointly. Harmonious Attention [18] is a combination of multiple attention mechanisms, such as Channel Attention [12] and Spatial Attention [29]. Inspired by Harmonious Attention, we propose to combine both types of partitions to form a spatial-channel partition. Comparison results between only spatial partition, only channel partition and spatial-channel partition are reported in Table 2. The performance of channel partition is better than that of spatial partition, because semantic concept-to-concept matching is more robust to misalignment than body part-to part matching. Spatial-channel partition can further enhance the performance.

Both partition and attention mechanism aim at enhancing the ability of neural networks to extract more discriminative features, but in opposite ways. Attention mechanism guides neural networks in locating the most important region in an image. As a consequence, secondary information may be neglected by attention mechanism. On the contrary, training local experts on partitioned parts enables neural networks to learn more local features. Heat maps of attention models in Figure 2 (g) and (f) keep less secondary information in the feature map than a ResNet-50 in Figure 2 (a), while partition based models in Figure 2 (b) to (e) keep more secondary information. Results in Table 1 validate that keeping secondary information by spatial-channel partition brings more improvement compared to removing secondary information by attention.

Model	Rank1	mAP
ResNet-50	89.5	73.3
ResNet-50 + channel attention (SENet)	90.8	75.6
ResNet-50 + spatial-channel partition	94.4	85.8

Table 1. Comparison of results (%) between attention and partition on Market-1501 dataset. SENet refers to Squeeze-and-Excitation Network [12]. Spatial-channel partition refers to the model trained with 2 spatial parts and 2 channel groups.

3. Proposed Framework

3.1. Spatial and Channel Partition Representation Network (SCR)

The general architecture of our proposed SCR network is represented in Figure 3. A batch of input images are fed into a backbone network. Last layers of backbone network are split into 3 independent branches in order to satisfy the need for a pyramidal structure, which generate 3 feature maps of equal size. A global feature map is extracted from each branch. The second and third feature maps are then partitioned into 2 and 3 spatial-channel parts respectively. A global max pooling (GMP) is used to replace global average pooling (GAP) in order to extract the most discriminative features in each part. Global and local feature maps are transferred to vectors with distinct dimensions. Next, dimensions of feature vectors are unified by 1*1 convolutional layers into 256. We train 13 fully connected layers as classifiers with 13 softmax cross-entropy losses respectively on each feature vector and 3 triplet losses on the 3 global feature vectors. More details are given in the following.

Backbone Network. Our proposed framework can take any convolutional neural network designed for image classification as backbone network, such as VGG [26] and ResNet [9]. To conduct a fair comparison, we follow previous state-of-the-art methods [30, 27] and use a ResNet-50 as our backbone. Two modifications are conducted: (1) the down-sampling with stride-2 convolution is replaced by a stride-1 convolution in the conv5_1 layer. (2) all the layers after conv4_1 layer are duplicated to form 3 independent branches. With these modifications, more high level features can be kept in the feature map.

Multiple Spatial-channel Partitions in a Pyramidal structure. To take full advantage of information contained in the feature map, global, spatial and channel partitioned features should be trained separately in the network. A pyramidal structure has proven to be beneficial for part based models in the previous state-of-the-art [37]. Thus, the second feature map is partitioned equally into 2 spatial parts and 2 channel groups. Similarly, the third feature map is partitioned into 3 spatial parts and 3 channel groups. With GMP, each partitioned map is transformed to a vector. Besides these local feature vectors, a global feature vector is extracted from each unpartitioned feature map. In total, there are 3 global vectors and 10 local vectors.

3.2. Loss Functions

Softmax Cross-Entropy loss. The Softmax Cross-Entropy loss in a mini-batch can be described as:

$$L_{CE} = -\sum_{i=1}^{N_i} \log\left(\frac{\exp(x[y])}{\sum_{j=1}^{N_{id}} \exp(x[j])}\right) \qquad (1)$$

where N_i denotes the number of images in the mini-batch, N_{id} is the number of identities in the whole training set. y is the ground truth identity of input image and $x[j]$ represents the output of fully-connected layer for jth identity.

Triplet loss. For a better performance on hard samples, the variant Batch Hard [10] is adopted. In a mini-batch

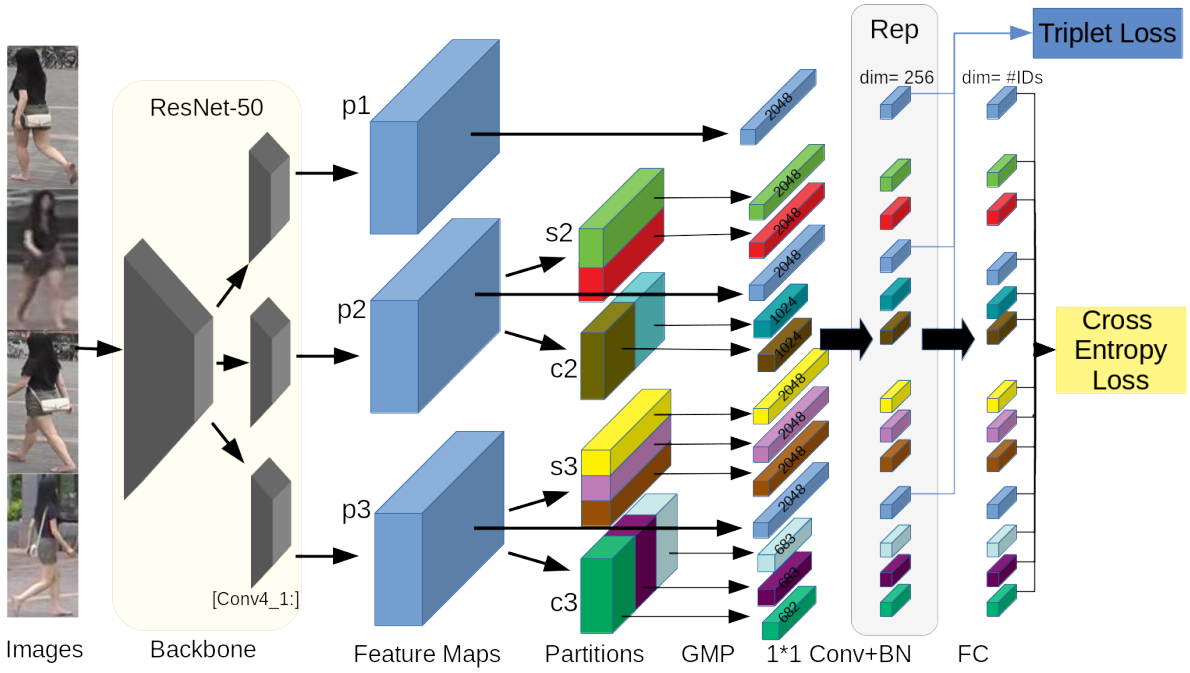

Figure 3. Spatial and Channel Partition Representation network. For the backbone network, we duplicate layers after conv4_1 into 3 identical but independent branches that generate 3 feature maps "p1", "p2" and "p3". Then, multiple spatial-channel partitions are conducted on the feature maps. "s2" and "c2" refer to 2 spatial parts and 2 channel groups. "s3" and "c3" refer to 3 spatial parts and 3 channel groups. After global max pooling (GMP), dimensions of global (dim = 2048) and local (dim = 2048, 1024*2 and 683*2+682) features are unified by 1*1 convolution (1*1 Conv) and batch normalization (BN) to 256. Then, fully connected layers (FC) give identity predictions of input images. All the dimension unified feature vectors (dim = 256) are aggregated together as appearance representation (Rep) for testing.

which contains P identities and K images for each identity, Batch Hard triplet loss aims at pulling the hardest positive pair (a, p) together while pushing the hardest negative pair (a, n) away by a margin. A Batch Hard triplet loss can be defined as:

$$L_{triplet} = \sum_{i=1}^{P}\sum_{a=1}^{K}[\max_{p=1,\ldots,K} \|\mathbf{a_i} - \mathbf{p_i}\|_2 - \min_{\substack{n=1,\ldots,K \\ j=1,\ldots,P \\ j\neq i}} \|\mathbf{a_i} - \mathbf{n_j}\|_2 + \alpha]_+ \quad (2)$$

where $\mathbf{a_i}$, $\mathbf{p_i}$ and $\mathbf{n_i}$ are the feature vectors of anchor, positive and negative samples respectively, and α is the margin to control the distance between positive and negative pair.

Total loss. Training the SCR model on global and local features jointly helps to build more robust representations. Local features extracted from small parts are sensitive to misalignment and viewpoint changes. Searching for the hardest positive and negative pairs with local features can be challenging, for example, we can not only look at the upper body when two people wear similar white T-shirts. Thus, the triplet loss is only employed on global features. Softmax cross-entropy loss helps to estimate the presence of specific features in small parts, which makes it more suitable for local features.

$$L_{total} = \lambda \frac{1}{N_{CE}} \sum_{i=1}^{N_{CE}} L_{CE} + \frac{1}{N_{triplet}} \sum_{i=1}^{N_{triplet}} L_{triplet} \quad (3)$$

where N_{CE} and $N_{triplet}$ are the number of softmax cross entropy losses and triplet losses respectively. In the SCR model, we have $N_{CE} = 13$ and $N_{triplet} = 3$. Parameter λ balances the contribution of two types of loss functions. Several possibilities of λ are tested in the next section to find an optimal setting for all experiments.

4. Experiments

4.1. Implementation Details

First of all, input images are resized to $384 * 192$. For the backbone network, we use a ResNet-50 pretrained on ImageNet [2] to accelerate the training process. All the layers after conv4_1 are duplicated into 3 independent branches. Each 1*1 convolutional layer is followed by a Batch Normalization [13] layer and a fully connected layer. These layers do not share weights. Following previous state-of-the-art methods [30, 27], we apply a standard Random Horizontal Flip for data augmentation. The batch size is set to

32 with randomly selected 8 identities and 4 images for each identity. We train our model with an Adam optimizer with AMSGrad setting [23] for 500 epochs. The weight decay factor for L2 regularization is set to 5e-4. The initial learning rate is set to 2e-4 and decay to 2e-5 after 300 epochs and to 2e-6 after 400 epochs. The margin α in triplet loss is set to 1.2 in all experiments and the parameter λ in total loss is set to 2. For the evaluation, we concatenate all the feature vectors after Batch Normalization layer together as the appearance representation for images in query and gallery sets. Our model is implemented on PyTorch framework and takes about 6 hours on a single NVIDIA 1080 Ti GPU for training on Market-1501 dataset.

4.2. Datasets and Protocols

To validate the effectiveness of our proposed SCR model, experiments are conducted on four mainstream Re-ID datasets: Market-1501 [39], DukeMTMC-reID [24, 41], CUHK03 [17] and MARS [38].

Image based datasets. Market-1501 dataset is collected in front of a supermarket in Tsinghua University. It contains 19,732 images of 751 identities in the training set and 12,936 images of 750 identities in the testing set. There are 17.2 images per identity in the training set. DukeMTMC-reID is a subset of the DukeMTMC dataset. It contains 16,522 images of 702 persons in the training set and 2,228 query images and 17,661 gallery images of 702 persons for testing. There are 23.5 images per identity in the training set. CUHK03 contains 14,096 images of 1,467 identities captured from Chinese University of Hong Kong campus. Each identity is captured from two cameras and has an average of 4.8 images in each camera. CUHK03 dataset provides both manually labeled bounding boxes and DPM [6] detected bounding boxes.

Video based dataset. MARS is an extension of the Market-1501 dataset. There are 509,914 bounding boxes for training, belonging to 8,298 tracklets of 625 identities. There are 681,089 bounding boxes for test (gallery+query), belonging to 12,180 tracklets of 636 identities.

Evaluation Protocols. Both Cumulative Matching Characteristics (CMC) and mean Average Precisions (mAP) are used in our experiments. CMC represents the matching accuracy of Person Re-ID and CMC at Rank1 is the most intuitive metric where each query has only one ground truth match. mAP is more appropriate for the case where each query has multiple gallery matches. On CUHK03 dataset, to simplify the evaluation procedure and meanwhile enhance the accuracy of the performance reflected by results, we employed the new protocol described in [42]. For MARS dataset, we conduct a tracklet-to-tracklet search by building an overall appearance representation on each tracklet instead of on single image. Re-ranking algorithm is not used to further improve mAP in all experiments.

Partition Type	CUHK03			
	Labelled		Detected	
	Rank1	mAP	Rank1	mAP
Spatial p1+p2(s2)+p3(s3)	75.9	72.1	75.6	71.8
Channel p1+p2(c2)+p3(c3)	81.5	77.4	77.9	73.9
Spatial-Channel p1+p2(s2c2)+p3(s3c3)	**83.8**	**80.4**	**82.2**	**77.6**

Table 2. Performance comparison (%) of different partition types (spatial partition, channel partition and spatial-channel partition) on CUHK03 dataset using the new protocol [42] where the bold font denotes the best partition type. "s2" and "s3" refer that the entire feature map is partitioned into 2 and 3 spatial parts, while "c2" and "c3" refer respectively to 2 and 3 channel groups.

4.3. Ablation Studies

To verify the effectiveness of each component in SCR and design an optimal architecture, we conduct extensive ablation studies on Market-1501, DukeMTMC-reID, CUHK03 and MARS datasets.

Partition Strategies. We conduct extensive experiments to validate the effectiveness of spatial-channel partition by comparing our proposed model with only spatial partitions, with only channel partitions and with spatial-channel partitions. These partition strategies are compared on the most challenging dataset CUHK03. Results are reported in Table 2. The model with spatial-channel partitions outperforms respectively the one with only channel partitions and the one with only spatial partitions by an average margin of 3% and 7% on CUHK03 dataset.

Pyramidal Multi-Branch Architectures. Each branch "p1", "p2" and "p3" is separately tested. As shown in Table 3, performances of "p2" and "p3" with spatial-channel partition have a significant improvement as compared to "p1" without partition. But the results are still below those of state-of-the-art. Thus, we adapt a pyramidal multi-branch architecture to our proposed SCR, which gives a boost to the performance of our model. We gradually increase the number of branch and report their performance in Table 3. Two phenomena are observed: 1) Spatial-channel partitions significantly increase the performance of the neural network in Re-ID task. 2) Multi-branch structure further enhances performance of the model.

Parameters in Total Loss. To balance contributions of softmax cross-entropy and triplet losses, a weight parameter λ should be determined. Four possibilities $\lambda = 1, 2, 3$ and without triplet loss are tested on CUHK03 dataset with both labelled and detected bounding boxes. Results in Table 4 shows that SCR gets best performance with $\lambda = 2$ on detected bounding boxes, while it gets best performance

Architecture	Number of Branches	Representation Dimension	Market-1501		DukeMTMC-reID		CUHK03-detected	
			Rank1	mAP	Rank1	mAP	Rank1	mAP
p1	1	256*1	89.5	73.3	83.3	66.8	52.4	46.3
p2(s2c2)	1	256*5	94.4	85.8	89.6	77.6	73.7	68.4
p3(s3c3)	1	256*7	94.5	86.3	89.2	78.7	75.1	70.7
p1+p2(s2c2)	2	256*(1+5)	94.9	87.3	89.5	78.6	74.8	70.2
p1+p2(s2c2)+p3(s3c3)	3	256*(1+5+7)	**95.7**	**89.0**	**91.1**	**81.4**	**82.2**	**77.6**

Table 3. Performance comparison (%) of the proposed SCR with different number of branches where the bold font denotes the best architecture. "p1", "p2" and "p3" refer to 3 feature maps in SCR. "s" and "c" represent "spatial" and "channel" respectively, followed by the number of parts. For instance, "s2" and "c2" refer that the entire feature map is partitioned into 2 spatial parts and 2 channel parts.

Loss Function	CUHK03			
	Labelled		Detected	
	Rank1	mAP	Rank1	mAP
w/o $L_{triplet}$	76.9	73.5	75.1	70.5
$\lambda = 1$	**84.8**	**81.4**	79.5	75.5
$\lambda = 2$	83.8	80.4	**82.2**	**77.6**
$\lambda = 3$	82.2	78.8	80.7	76.8

Table 4. Performance comparison (%) of training SCR with different parameter values for λ from L_{total}. The bold font denotes the best parameter.

Temporal Pooling	MARS	
	Rank1	mAP
TP(R)	84.5	78.9
TPP(R)	86.6	80.8
TP(E)	85.7	79.5
TPP(E)	**87.3**	**81.3**

Table 5. Comparison of different temporal pooling strategies where the bold font denotes the best method. "R" and "E" refer respectively to Random Sampling and Even Sampling. "TP" refers to conventional Temporal Pooling. "TPP" refers to Temporal Partiton Pooling.

with $\lambda = 1$ on labeled bounding boxes. To form a unified framework, we set $\lambda = 2$ for all experiments.

Temporal Partition Pooling (TPP). For video based Re-ID, a traditional approach for building a tracklet representation is to use a temporal average (or max) pooling on all sampled image representations for the tracklet. To generalize partition strategies for the video-based Re-ID task, we conduct a partition on the temporal dimension over the tracklet. Instead of adopting directly a temporal pooling on all sampled images in a tracklet, we firstly split the images into several sub-tracklets and use the temporal pooling separately on each sub-tracklet. Representations of sub-tracklets are concatenated together to form a final representation of the tracklet. To validate the performance of our proposed TPP, we fix the sample size to 15 and partition the 15 images into 3 groups (beginning, middle, end). Different sample size and group number are tested but they do not have a strong effect on results. A temporal average pooling is performed on each sub-tracklet. Results in Table 5 show that temporal partition can enhance the performance of our model for the video-based Re-ID task.

Method	Market-1501		DukeMTMC-reID	
	Rank1	mAP	Rank1	mAP
HA-CNN [18]	91.2	75.7	80.5	63.8
Mancs [28]	93.1	82.3	84.9	71.8
PCB+RPP [27]	93.8	81.6	83.3	69.2
SCPNet-a [4]	94.1	81.8	84.4	68.5
HPM [7]	94.2	82.7	86.6	74.3
CAMA [35]	94.7	84.5	85.8	72.9
MGN [30]	**95.7**	86.9	88.7	78.4
CPM [37]	**95.7**	88.2	89.0	79.0
SCR(ours)	**95.7**	**89.0**	**91.1**	**81.4**

Table 6. Comparison of supervised results (%) on Market-1501 and DukeMTMC-reID dataset.

4.4. Comparison with State-of-the-art

We compare our proposed model SCR with current state-of-the-art methods on the 4 candidate datasets.

Results on Market-1501. Comparisons between SCR and state-of-the-art methods on Market-1501 are shown in Table 6. Our proposed SCR achieves a mAP of 89.0% under single query setting, which surpasses the previous most performant model CPM [37] by 0.8%. To get a better understanding on how our proposed SCR can outperform previous state-of-the-art, we compare some retrieved results between PCB [27] and our SCR in Figure 4. For the first query image, pose similarity leads to mismatch in PCB, while our SCR succeeds to consider some high level semantic features neglected in PCB, *e.g.*, color of shorts. For the second query image, PCB are prone to error when person image is misaligned. These results confirm the effectiveness of spatial-channel partition on keeping more salient information and that of pyramidal structure to deal with misalignment.

Results on DukeMTMC-reID. Results of SCR and pre-

Method	CUHK03			
	Labelled		Detected	
	Rank1	mAP	Rank1	mAP
HA-CNN [18]	44.4	41.0	41.7	38.6
PCB+RPP [27]	-	-	63.7	57.5
HPM [7]	-	-	63.9	57.5
MGN [30]	68.0	67.4	68.0	66.0
CAMA [35]	70.1	66.5	66.6	64.2
CPM [37]	78.9	76.9	78.9	74.8
SCR(ours)	**83.8**	**80.4**	**82.2**	**77.6**

Table 7. Comparison of supervised results (%) on CUHK03 dataset using the new protocol [42] .

Method	MARS	
	Rank1	mAP
IDE+Kissme [38]	68.3	49.3
TriNet [10]	79.8	67.7
DRSTA [16]	82.3	65.8
M3D [15]	84.4	74.0
SCR(ours)	**87.3**	**81.3**

Table 8. Comparison of supervised results (%) on MARS dataset.

Method	M $\rightarrow$ D		D $\rightarrow$ M	
	Rank1	mAP	Rank1	mAP
SPGAN [3]	41.1	22.3	51.5	22.8
TJ-AIDL [31]	44.3	23.0	58.2	26.5
ATNet [21]	45.1	24.9	55.7	25.6
HHL [43]	46.9	27.2	**62.2**	**31.4**
SCR(ours)	**53.6**	**32.4**	59.7	30.6

Table 9. Comparison of unsupervised cross-domain results (%). M $\rightarrow$ D refers to training on Market-1501 and testing on DukeMTMC-reID. D $\rightarrow$ M refers to training on DukeMTMC-reID and testing on Market-1501.

vious state-of-the-art methods on DukeMTMC-reID dataset are reported in Table 6. This dataset is more chanlleng-ing than Market-1501 because it has 8 different cameras and bounding box size varies dramatically across different camera views. Our SCR network also performs excellently on DukeMTMC-reID dataset. SCR outperforms the former state-of-the-art by 2.1% on Rank1 and 2.4% on mAP.

Results on CUHK03. Table 7 shows results on CUHK03 dataset. Due to less training samples per iden-tity, algorithms tend to get lower scores on CUHK03, which makes CUHK03 the most challenging evaluation protocol. With the same parameter settings, SCR outperforms previ-ous state-of-the-art CPM by a large margin.

Results on MARS. To validate the adaptability of our model in the video-based Re-ID task, we conduct experi-ments on MARS dataset and report results in Table 8. Our

Figure 4. Examples of several mismatched samples in PCB [27] on Market-1501 dataset, which are addressed by our proposed SCR. Red borders refers to mismatched samples. "#1", "#2" and "#3" correspond to top 3 retrieved gallery samples.

model is able to outperform current state-of-the-art video-based models. SCR outperforms the previous most perfor-mant model M3D [15] by a large margin.

Unsupervised cross-domain results. Our proposed method also shows a strong generalizability on unsuper-vised cross-domain problem, in which a model is trained on a source domain and tested on a target domain. We compare results of SCR and unsupervised cross-domain methods in Table 9. Without using unlabeled images in target domain like [3, 43, 21] or extra attribute annotation [31], our SCR outperforms previous state-of-the-art under a direct deploy-ment (no re-training on target domain) setting.

5. Conclusion

In this work, we carry out a comparative study between spatial, channel partitions and attention mechanism. Based on this study, a novel end-to-end trainable Spatial and Chan-nel partition Representation network (SCR) is proposed to maintain salient local information by spatial-channel parti-tions. By combining spatial-channel partitioned local fea-tures with global features, our SCR model is able to build a discriminative and generalizable representation for each sample for Re-ID task. In addition, to address the misalign-ment problem, we use spatial-channel partitions in a pyra-midal multi-branch architecture, which can further improve the robustness of local features. To get a better performance in video based Re-ID, partition is extended to the temporal dimension. The effectiveness of each proposed component is validated in the ablation studies. Crucial components, like spatial-channel partition and temporal partition pooling, can be easily embedded into other part based models for Re-ID. By incorporating all these components, our well-designed method outperforms current state-of-the-art in both image and video based supervised Re-ID task, as well as in unsu-pervised cross domain task.

References

[1] S. Bak, E. Corvée, F. Brémond, and M. Thonnat. Person re-identification using spatial covariance regions of human body parts. *2010 7th IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 435–440, 2010.

[2] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[3] W. Deng, L. Zheng, G. Kang, Y. Yang, Q. Ye, and J. Jiao. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 994–1003, 2017.

[4] X. Fan, H. Luo, X. Zhang, L. He, C. Zhang, and W. Jiang. Scpnet: Spatial-channel parallelism network for joint holistic and partial person re-identification. In *ACCV*, 2018.

[5] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. Person re-identification by symmetry-driven accumulation of local features. *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2360–2367, 2010.

[6] P. F. Felzenszwalb, D. A. McAllester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model. *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008.

[7] Y. Fu, Y. Wei, Y. Zhou, H. Shi, G. Huang, X. Wang, Z. Yao, and T. S. Huang. Horizontal pyramid matching for person re-identification. In *AAAI*, 2019.

[8] D. Gray and H. Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *ECCV*, 2008.

[9] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[10] A. Hermans, L. Beyer, and B. Leibe. In defense of the triplet loss for person re-identification. *CoRR*, abs/1703.07737, 2017.

[11] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *CoRR*, abs/1704.04861, 2017.

[12] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In *CVPR*, 2018.

[13] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML 2015*, 2015.

[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Commun. ACM*, 60:84–90, 2012.

[15] J. Li, S. Zhang, and T. Huang. Multi-scale 3d convolution network for video based person re-identification. In *AAAI*, 2019.

[16] S. Li, S. Bak, P. Carr, and X. Wang. Diversity regularized spatiotemporal attention for video-based person re-identification. In *CVPR*, 2018.

[17] W. Li, R. Zhao, T. Xiao, and X. Wang. Deepreid: Deep filter pairing neural network for person re-identification. *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 152–159, 2014.

[18] W. Li, X. Zhu, and S. Gong. Harmonious attention network for person re-identification. In *CVPR*, 2018.

[19] S. Liao, Y. Hu, X. Zhu, and S. Z. Li. Person re-identification by local maximal occurrence representation and metric learning. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2197–2206, 2015.

[20] S. Lin, H. Li, V. Sanchez, and A. C. Kot. Multi-task mid-level feature alignment network for unsupervised cross-dataset person re-identification. In *BMVC*, 2018.

[21] J. Liu, Z.-J. Zha, D. Chen, R. Hong, and M. Wang. Adaptive transfer network for cross-domain person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[22] A. Mignon and F. Jurie. Pcca: A new approach for distance learning from sparse pairwise constraints. *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2666–2672, 2012.

[23] S. J. Reddi, S. Kale, and S. Kumar. On the convergence of adam and beyond. In *International Conference on Learning Representations*, 2018.

[24] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *European Conference on Computer Vision workshop on Benchmarking Multi-Target Tracking*, 2016.

[25] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017.

[26] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[27] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *ECCV*, 2018.

[28] C. Wang, Q. Zhang, C. Huang, W. Liu, and X. Wang. Mancs: A multi-task attentional network with curriculum sampling for person re-identification. In *ECCV*, 2018.

[29] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang. Residual attention network for image classification. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6450–6458, 2017.

[30] G. Wang, Y. Yuan, X. Chen, J. Li, and X. Zhou. Learning discriminative features with multiple granularities for person re-identification. In *ACM Multimedia*, 2018.

[31] J. Wang, X. Zhu, S. Gong, and W. Li. Transferable joint attribute-identity deep learning for unsupervised person re-identification. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2275–2284, 2018.

[32] Y. Wu and K. He. Group normalization. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[33] T. Xiao, H. Li, W. Ouyang, and X. Wang. Learning deep feature representations with domain guided dropout for person re-identification. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1249–1258, 2016.

[34] S. Xie, R. B. Girshick, P. Dollár, Z. Tu, and K. He. Aggregated residual transformations for deep neural networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5987–5995, 2017.

[35] W. Yang, H. Huang, Z. Zhang, X. Chen, K. Huang, and S. Zhang. Towards rich feature discovery with class activation maps augmentation for person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[36] H. Yao, S. Zhang, R. Hong, Y. Zhang, C. Xu, and Q. Tian. Deep representation learning with part loss for person re-identification. *IEEE Transactions on Image Processing*, 2019.

[37] F. Zheng, C. Deng, X. Sun, X. Jiang, X. Guo, Z. Yu, F. Huang, and R. Ji. Pyramidal person re-identification via multi-loss dynamic training. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[38] L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang, and Q. Tian. Mars: A video benchmark for large-scale person re-identification. In *ECCV 2016*, 2016.

[39] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person re-identification: A benchmark. *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1116–1124, 2015.

[40] Z. Zheng, L. Zheng, and Y. Yang. A discriminatively learned cnn embedding for person reidentification. *TOMCCAP*, 14:13:1–13:20, 2017.

[41] Z. Zheng, L. Zheng, and Y. Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 3774–3782, 2017.

[42] Z. Zhong, L. Zheng, D. Cao, and S. Li. Re-ranking person re-identification with k-reciprocal encoding. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3652–3661, 2017.

[43] Z. Zhong, L. Zheng, S. Li, and Y. Yang. Generalizing a person retrieval model hetero- and homogeneously. In *The European Conference on Computer Vision (ECCV)*, September 2018.