Person Re-Identification by Deep Joint Learning of Multi-Loss Classification

Wei Li, Xiatian Zhu, Shaogang Gong
Queen Mary University of London, London E1 4NS, UK
{w.li, xiatian.zhu, s.gong}@qmul.ac.uk

Abstract

Existing person re-identification (re-id) methods rely mostly on either localised or global feature representation *alone*. This ignores their joint benefit and mutual complementary effects. In this work, we show the advantages of jointly learning local and global features in a Convolutional Neural Network (CNN) by aiming to discover correlated local and global features in different context. Specifically, we formulate a method for joint learning of local and global feature selection losses designed to optimise person re-id when using *only* generic matching metrics such as the L2 distance. We design a novel CNN architecture for Jointly Learning Multi-Loss (JLML). Extensive comparative evaluations demonstrate the advantages of this new JLML model for person re-id over a wide range of state-of-the-art re-id methods on four benchmarks (VIPeR, GRID, CUHK03, Market-1501).

1 Introduction

Person re-identification (re-id) is about matching identity classes in detected person bounding box images from non-overlapping camera views over distributed open spaces. This is an inherently challenging task because person visual appearance may change dramatically in different camera views due to unknown changes in illumination, occlusion, and background clutter [Gong *et al.*, 2014]. Existing re-id studies typically focus on either feature representation [Gray and Tao, 2008; Farenzena *et al.*, 2010; Kviatkovsky *et al.*, 2013; Zhao *et al.*, 2013; Liao *et al.*, 2015; Matsukawa *et al.*, 2016a; Ma *et al.*, 2017] or matching distance metrics [Koestinger *et al.*, 2012; Xiong *et al.*, 2014; Zheng *et al.*, 2013; Paisitkriangkrai *et al.*, 2015; Zhang *et al.*, 2016; Wang *et al.*, 2014b; 2016b; 2016c; Chen *et al.*, 2017b] or their combination in deep learning framework [Li *et al.*, 2014; Ahmed *et al.*, 2015; Wang *et al.*, 2016a; Xiao *et al.*, 2016; Subramaniam *et al.*, 2016; Chen *et al.*, 2017a]. Regardless, the overall objective is to obtain a view- and location-invariant representation. We consider that learning any matching distance metric is intrinsically learning a global feature transformation across domains (two disjoint cameras) thus obtaining a "normalised" representation for matching.

Most re-id features are typically hand-crafted to encode *local* topological and/or spatial structural information, by different image decomposition schemes such as horizontal stripes [Gray and Tao, 2008; Kviatkovsky *et al.*, 2013], body parts [Farenzena *et al.*, 2010], and patches [Zhao *et al.*, 2013; Matsukawa *et al.*, 2016a; Liao *et al.*, 2015]. These localised features are effective for mitigating the person pose and detection misalignment in re-id matching. More recent deep re-id models [Xiao *et al.*, 2016; Wang *et al.*, 2016a; Chen *et al.*, 2017a; Ahmed *et al.*, 2015] benefit from the availability of larger scale datasets such as CUHK03 [Li *et al.*, 2014] and Market-1501 [Zheng *et al.*, 2015] and from lessons learned on other vision tasks [Krizhevsky *et al.*, 2012; Girshick *et al.*, 2014]. In contrast to *local* hand-crafted features, deep models, in particular Convolutional Neural Networks (CNN) [LeCun *et al.*, 1998], favour intrinsically in learning *global* feature representations with a few exceptions. They have been shown to be effective for re-id.

We consider that either local or global feature learning *alone* is suboptimal. This is motivated by the human visual system that leverages both global (contextual) and local (saliency) information concurrently [Navon, 1977; Torralba *et al.*, 2006]. This intuition for *joint learning* aims to extract correlated complementary information in different context whilst *satisfying the same learning constraint* therefore achieving more reliable recognition. To that end, we need to address a number of non-trivial problems: (i) the model learning behaviour in satisfying the same label constraint may be different at the local and global levels; (ii) any complementary correlation between local and global features is unknown and may vary among individual instances, therefore must be learned and optimised consistently across data; (iii) people's appearance in public scenes is diverse in both pattens and configurations. This makes it challenging to learn correlations between local and global features *for all appearances*.

This work aims to formulate a deep learning model for jointly optimising local and global feature selections concurrently and to improve person re-id using *only* generic matching metrics such as the L2 distance. We explore a deep learning approach for its potential superiority in learning from large scale data [Xiao *et al.*, 2016; Chen *et al.*, 2017a]. For the bounding box image based person re-id, we consider the entire person in the image as a *global scene context* and body parts of the person as *local information sources*, both are

subject to the surrounding background clutter within an image, and potentially also misalignment and partial occlusion from poor detection. In this setting, we wish to discover and optimise jointly correlated complementary feature selections in the local and global representations, both subject to the same label constraint concurrently. Whilst the former aims to address detection misalignment and occlusion by localised fine-grained saliency information, the latter exploits holistic coarse-grained context for more robust global matching.

Our **contributions** are: **(I)** We propose the idea of learning concurrently both local and global feature selections for optimising feature discriminative capabilities in different context whilst performing the same person re-id tasks. This is currently under-studied in the person re-id literature to our best knowledge. **(II)** We formulate a novel *Joint Learning Multi-Loss* (JLML) CNN model for not only learning both global and local discriminative features in different context by optimising multiple classification losses on the same person label information concurrently, but also utilising their complementary advantages jointly in coping with local misalignment and optimising holistic matching criteria for person re-id. This is achieved with a deep two-branch CNN architecture by imposing inter-branch interaction between the local and global branches, and enforcing a separate learning objective loss function to each branch for learning independent discriminative capabilities. **(III)** We introduce a structured sparsity based feature selection learning mechanism for improving multi-loss joint feature learning robustness w.r.t. noise and data covariance between local and global representations. Extensive evaluations demonstrate the superiority of the proposed JLML model over a wide range of existing state-of-the-art re-id models on four benchmark datasets.

Related Works. The JLML method is related to the saliency learning based models [Zhao *et al.*, 2013; Wang *et al.*, 2014a] in terms of modelling localised part importance. However, these existing methods consider only the patch appearance statistics within individual locations but no global feature representation learning, let alone the correlation and complementary information discovery between local and global features as modelled by the JLML. Whilst the more recent SCS [Chen *et al.*, 2016] and MCP [Cheng *et al.*, 2016] consider both levels of representation, the JLML model differs significantly from them: **(i)** The SCS method focuses on supervised metric learning, whilst the JLML aims at joint discriminative feature learning and needs only generic metrics for re-id matching. **(ii)** The local and global branches of the MCP model are supervised and optimised by a triplet ranking loss, in contrast to the proposed multiple classification loss design. **(iii)** The JLML is uniquely capable of performing structured feature sparsity regularisation.

2 Model Design

2.1 Problem Definition

We assume a set of n training images $\mathcal{I} = \{\boldsymbol{I}_i\}_{i=1}^n$ with the corresponding identity labels as $\mathcal{Y} = \{y_i\}_{i=1}^n$. These training images capture the visual appearance of n_{id} (where $y_i \in [1, \cdots, n_{\text{id}}]$) different people under non-overlapping camera views. We formulate a Joint Learning Multi-Loss (JLML) CNN model that aims to discover and capture concurrently complementary discriminative information about a person image from both local and global visual features of the image in order to optimise person re-id under significant viewing condition changes across locations. This is in contrast to most existing re-id methods typically depending only on either local or global features alone.

2.2 Joint Learning Multi-Loss

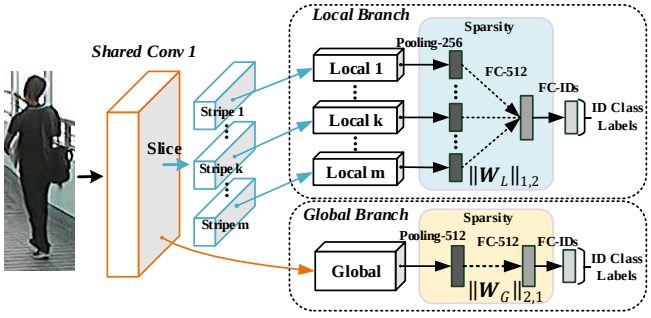

Figure 1: The Joint Learning Multi-Loss (JLML) CNN model.

The overall design of the proposed JLML model is depicted in Figure 1. This JLML model consists of a two-branches CNN network: (1) One *local branch* of m streams of an identical structure with each stream learning the most discriminative local visual features for one of m local image regions of a person bounding box image; (2) Another *global branch* responsible for learning the most discriminative global level features from the entire person image. For concurrently optimising per-branch discriminative feature representations and discovering correlated complementary information between local and global feature selections, a *joint learning* scheme that subjects both local and global branches to the same identity label supervision is formulated with two underlying principles:

(I) Shared low-level features. We construct the global and local branches on a shared lower conv layer, in particular the first conv layer, for facilitating inter-branch common learning. The intuition is that, the lower conv layers capture low-level features such as edges and corners which are common to all patterns in the same images. This shared learning is similar in spirit to multi-task learning [Argyriou *et al.*, 2007], where the local and global feature learning branches are two related learning tasks. Sharing the low-level conv layer reduces the model parameter size therefore model overfitting risks. This is especially critical in learning person re-id models when labelled training data is limited.

(II) Multi-task independent learning. To maximise the learning of complementary discriminative features from local and global representations, the remaining layers of the two branches are learned independently subject to given identity labels. That is, the JLML model aims to learn concurrently multiple identity feature representations for different local image regions and the entire image, all of which aim to maximise the *same* identity matching *both* individually and collectively at the same time. Independent multi-task learning aims to preserve both local saliency in feature selection and

global robustness in image representation. To that end, the JLML model is designed to perform *multi-task independent learning subject to shared identity label constraints* by allocating each branch with a separate objective loss function. By doing so, the per-branch learning behaviour is conditioned independently on the respective feature representation. We call this branch-wise loss formulation as the **MultiLoss** design.

Table 1: JLML-ResNet39 model (1.5 billion FLOPs). MP: Max-Pooling; AP: Average-Pooling; S: Stride; SL: Slice; CA: Concatenation; G: Global; L: Local.

Layer #	Layer	Output Size	Global Branch	Local Branch
1	conv1	112×112	3×3, 32, S-2	
9	conv2_x	G: 56×56 L: 28×56	3×3 MP, S-2 $\begin{bmatrix}1\times1, 32\\3\times3, 32\\1\times1, 64\end{bmatrix}\times3$	SL-4, 2×2 MP, S-1 $\begin{bmatrix}1\times1, 16\\3\times3, 16\\1\times1, 32\end{bmatrix}\times3$
9	conv3_x	G: 28×28 L: 14×28	$\begin{bmatrix}1\times1, 64\\3\times3, 64\\1\times1, 128\end{bmatrix}\times3$	$\begin{bmatrix}1\times1, 32\\3\times3, 32\\1\times1, 64\end{bmatrix}\times3$
9	conv4_x	G: 14×14 L: 7×14	$\begin{bmatrix}1\times1, 128\\3\times3, 128\\1\times1, 256\end{bmatrix}\times3$	$\begin{bmatrix}1\times1, 64\\3\times3, 64\\1\times1, 128\end{bmatrix}\times3$
9	conv5_x	G: 7×7 L: 4×7	$\begin{bmatrix}1\times1, 256\\3\times3, 256\\1\times1, 512\end{bmatrix}\times3$	$\begin{bmatrix}1\times1, 128\\3\times3, 128\\1\times1, 256\end{bmatrix}\times3$
1	fc	1×1	7×7 AP $\begin{bmatrix}1\times1, 512\end{bmatrix}$	4×7 AP, CA-4 $\begin{bmatrix}1\times1, 512\end{bmatrix}$
1	fc	1×1	ID #	ID #

Network Construction. We adopt the Residual CNN unit [He *et al.*, 2016] as the JLML's building blocks due to its capacity for deeper model design whilst retaining a smaller model parameter size. Specifically, we customise the ResNet50 architecture in both layer and filter numbers and design the JLML model as a 39 layers 2-branches ResNet (**JLML-ResNet39**) tailored for re-id tasks. The configuration of JLML-ResNet39 is given in Table 1. Note that, the ReLU rectification non-linearity [Krizhevsky *et al.*, 2012] after each conv layer is omitted for brevity.

Feature Selection. To optimise JLML model learning robustness against noise and diverse data source, we introduce a feature selection capability in JLML by a structure sparsity induced regularization [Kong *et al.*, 2014; Wang *et al.*, 2013]. Our idea is to have a competing-to-survive mechanism in feature learning that discourages irrelevant features whilst encourages discriminative features concurrently in different local and global context to maximise a shared identity matching objective. To that end, we sparsify the global feature representation with a group LASSO [Wang *et al.*, 2013]:

$$\ell_{2,1} = \|\boldsymbol{W}_G\|_{2,1} = \sum_{i=1}^{d_g} \|\boldsymbol{w}_g^i\|_2 \tag{1}$$

where $\boldsymbol{W}_G = [\boldsymbol{w}_g^1, \cdots, \boldsymbol{w}_g^{d_g}] \in \mathcal{R}^{c_g \times d_g}$ is the parameter matrix of the global branch feature layer taking as input d_g dimensional vectors from the previous layer and outputting c_g dimensional (512-D) feature representation. Specifically, with the ℓ_1 norm applied on the ℓ_2 norm of $\boldsymbol{w}_g^i$, our aim is to learn selectively feature importance subject to both the sparsity principle and the identity label constraint simultaneously. Similarly, we also enforce a local feature sparsity constraint by an exclusive group LASSO [Kong *et al.*, 2014]:

$$\ell_{1,2} = \|\boldsymbol{W}_L\|_{1,2} = \sum_{i=1}^{c_l} \sum_{j=1}^{m} \|\boldsymbol{w}_{l,j}^i\|_1^2 \tag{2}$$

where $\boldsymbol{W}_L$ is the parameter matrix of the local branch feature layer with $m \times d_l$ and c_l (512) as the input and output dimensions (m the image stripe number), and $\boldsymbol{w}_{l,j}^i \in \mathcal{R}^{d_l \times 1}$ defines the parameter vector for contributing the i-th output feature dimension from the j-th local input feature vector, $j \in [1, 2, \cdots, m]$. This $\ell_{1,2}$ regularisor performs sparse feature selection for individual image regions in conjunction with the global feature selection learning.

Loss Function. For model training, we utilise the cross-entropy *classification* loss function for both global and local branches so to optimise person *identity classification* given training labels of multiple person classes extracted from pairwise labelled re-id dataset. Formally, we predict the posterior probability $\tilde{y}_i$ of image $\boldsymbol{I}_i$ over the given identity label y_i:

$$p(\tilde{y}_i = y_i | \boldsymbol{I}_i) = \frac{\exp(\boldsymbol{w}_{y_i}^\top \boldsymbol{x}_i)}{\sum_{k=1}^{|n_{\text{id}}|} \exp(\boldsymbol{w}_k^\top \boldsymbol{x}_i)} \tag{3}$$

where $\boldsymbol{x}_i$ refers to the feature vector of $\boldsymbol{I}_i$ from the corresponding branch, and $\boldsymbol{w}_k$ the prediction function parameter of training identity class k. The training loss on a batch of n_{bs} images is computed as:

$$l = -\frac{1}{n_{\text{bs}}} \sum_{i=1}^{n_{\text{bs}}} \log\Big(p(\tilde{y}_i = y_i | \boldsymbol{I}_i)\Big) \tag{4}$$

Combined with the group sparsity based feature selection regularisations, we have the final loss function for the global and local branch sub-networks as:

$$l_{\text{global}} = l + \lambda_{\text{global}} \|\boldsymbol{W}_G\|_{2,1}, \;\; l_{\text{local}} = l + \lambda_{\text{local}} \|\boldsymbol{W}_L\|_{1,2} \tag{5}$$

where λ_{global} and λ_{local} control the balance between the identity label loss and the feature selection sparsity regularisation. We empirically set $\lambda_{\text{local}} = \lambda_{\text{global}} = 5{\times}10^{-4}$.

Choice of Loss Function. Our JLML model learning deploys a *classification* loss function. This differs significantly from the *contrastive* loss functions used by most existing deep re-id methods designed to exploit pairwise re-id labels defined by *both* positive and negative pairs, such as the pairwise verification [Varior *et al.*, 2016; Subramaniam *et al.*, 2016; Ahmed *et al.*, 2015; Li *et al.*, 2014], triplet ranking [Cheng *et al.*, 2016], or both [Wang *et al.*, 2016a; Chen *et al.*, 2017a]. Our JLML model training does *not* use any labelled negative pairs inherent to all person re-id training data, and we extract identity class labels from only positive pairs. The motivations for our JLML classification loss based learning are: **(i)** Significantly *simplified* training data batch construction, e.g. random sampling with no notorious tricks required, as shown by other deep classification methods [Krizhevsky *et al.*, 2012]. This makes our JLML model more scalable in real-world applications with very large training population sizes when available and/or imbalanced training data sampling from different camera views. This also eliminates the

undesirable need for carefully forming pairs and/or triplets in preparing re-id training splits, as in most existing methods, due to the inherent imbalanced negative and positive pair size distributions. **(ii)** Visual psychophysical findings suggest that representations optimised for classification tasks generalise well to novel categories [Edelman, 1998]. We consider that re-id tasks are about model generalisation to unseen test identity classes given training data on *independent* seen identity classes. Our JLML model learning exploits this general classification learning principle beyond the strict pairwise relative verification loss in existing re-id models.

Model Training. We adopt the Stochastic Gradient Descent (SGD) optimisation algorithm [Krizhevsky *et al.*, 2012] to perform the batch-wise joint learning of local and global branches. Note that, with SGD we can naturally synchronise the optimisation processes of the two branches by constraining their learning behaviours subject to the same identity label information at each update. This is likely to avoid representation learning divergence between two branches and help enhance the correlated complementary learning capability.

2.3 Person Re-Id by Generic Distance Metrics

Once the JLML model is learned, we obtain a 1,024-D joint representation by concatenating the local (512-D) and global (512-D) feature vectors (the fc layers in Table 1). For person re-id, we deploy this 1,024-D deep feature representation using *only* a generic distance metric *without* camera-pair specific distance metric learning, e.g. the L2 distance.

3 Experiments

Datasets. For evaluation, we used four benchmarking re-id datasets, VIPeR [Gray and Tao, 2008], GRID [Loy *et al.*, 2009], CUHK03 [Li *et al.*, 2014], and Market-1501 [Zheng *et al.*, 2015]. These datasets present a wide range of re-id evaluation scenarios with different population sizes under different challenging viewing conditions (Figure 2 and Table 2).

(a) VIPeR (b) GRID (c) CUHK03 (d) Market

Figure 2: Example cross-view image pairs from four re-id datasets.

Table 2: Settings of person re-id datasets. TS: Test Setting; SS: Single-Shot; SQ: Single-Query; MQ: Multi-Query.

Dataset	Cams	IDs	Train IDs	Test IDs	Labelled	Detected	TS
VIPeR	2	632	316	316	1,264	0	SS
GRID	8	250	125	125	1,275	0	SS
CUHK03	6	1,467	1,367	100	14,097	14,097	SS
Market	6	1,501	751	750	0	32,668	SQ/MQ

Evaluation Protocol. We adopted the standard supervised re-id setting to evaluate the proposed JLML model (Sec. 3.1). The training and test data splits and the test settings of each dataset is given in Table 2. Specifically, on VIPeR, we split randomly the whole population (632 people) into two halves: One for training (316) and another for testing (316). We repeated 10 trials of random people splits and used the averaged results. On GRID, the training/test split was 125/125 with 775 distractor people included in the test gallery. We used the benchmarking 10 people splits [Loy *et al.*, 2009] and the averaged performance. On CUHK03, following [Li *et al.*, 2014] we repeated 20 times of random 1260/100 training/test splits and reported the averaged accuracies under the single-shot evaluation setting. On Market-1501, we used the standard training/test split (750/751) [Zheng *et al.*, 2015]. We used the cumulative matching characteristic (CMC) to measure re-id accuracy on all benchmarks, except on Market-1501 we also used the recall measure by mean Average Precision (mAP).

Table 3: Person re-id method categorisation by features and metrics. Cat: Category; DL: Deep Learning; CPSL: Camera-Pair Specific Learning; DVM: Deep Verification Metric; DVM, L2: Ensemble of DVM and L2; CHS: Fusion of Colour, HOG, SILPT features.

Cat	Method	Feature		Metric	
		Hand-Crafted	DL	CPSL	Generic
A	**XQDA** [Liao *et al.*, 2015]	LOMO	-	XQDA	-
	GOG [Matsukawa *et al.*, 2016b]	GOG	-	XQDA	-
	NFST [Zhang *et al.*, 2016]	LOMO, KCCA	-	NSFT	-
	SCS [Chen *et al.*, 2016]	CHS	-	SCS	-
B	**DCNN+** [Ahmed *et al.*, 2015]	-	DCNN+	DVM	-
	X-Corr [Subramaniam *et al.*, 2016]	-	X-Corr	DVM	-
	MTDnet [Chen *et al.*, 2017a]	-	MTDnet	DVM, L2	-
C	**S-CNN** [Varior *et al.*, 2016]	-	S-CNN	-	L2
	DGD [Xiao *et al.*, 2016]	-	DGD	-	L2
	MCP [Cheng *et al.*, 2016]	-	MCP	-	L2
	JLML (Ours)	-	JLML	-	L2

Competitors. We compared the JLML model against 10 existing state-of-the-art methods as listed in Table 3. They range from hand-crafted and deep learning features to domain-specific distance metric learning methods. We summarise them into three categories: **(A)** Hand-crafted (feature) with domain-specific distance learning (metric); **(B)** Deep learning (feature) with domain-specific deep verification (metric) learning; **(C)** Deep learning (feature) with generic non-learning L2 distance (metric).

Table 4: JLML training parameters. BLR: base learning rate; LRP: learning rate policy; MOT: momentum; IT: iteration; BS: batch size.

Parameter	BLR	LRP	MOT	IT #	BS
Pre-train	0.01	step (0.1, 100K)	0.9	300K	32
Train	0.01	step (0.1, 20K)	0.9	50K	32

Implementation. We used the Caffe framework [Jia *et al.*, 2014] for our JLML model implementation. We started by pre-training the JLML model on ImageNet (ILSVRC2012). Subsequently, for CUHK03 or Market, we used only their own training data for model fine-tuning, i.e. ImageNet $\rightarrow$ CUHK03/Market; For VIPeR or GRID, we pre-trained JLML on CUHK03+Market (whole datasets), and then fine-tuned on their respective training images, i.e. ImageNet $\rightarrow$ CUHK03+Market $\rightarrow$ VIPeR/GRID. All person images were resized to 224×224 in pixel. For local branch, according to a coarse body part layout we evenly decomposed the whole shared conv feature maps (i.e. the entire image) into four ($m = 4$) horizontal strip-regions. We used the same parameter settings (summarised in Table 4) for pre-training and training the JLML model on all datasets. We also adopted the stepped learning rate policy, e.g. dropping the learning rate by a factor of 10 every 100K iterations for JLML pre-training and every 20K iterations for JLML training. We utilised the L2 distance as the default matching metric.

Table 5: CUHK03 evaluation. $1^{st}/2^{nd}$ best in red/blue.

Cat	Annotation	Labelled				Detected			
	Rank (%)	R1	R5	R10	R20	R1	R5	R10	R20
A	XQDA	55.2	77.1	86.8	83.1	46.3	78.9	83.5	93.2
	GOG	67.3	91.0	96.0	-	65.5	88.4	93.7	-
	NSFT	62.5	90.0	94.8	98.1	54.7	84.7	94.8	95.2
B	DCNN+	54.7	86.5	93.9	98.1	44.9	76.0	83.5	93.2
	X-Corr	72.4	95.5	-	98.4	72.0	96.0	-	98.2
	MTDnet	74.7	96.0	97.5	-	-	-	-	-
C	S-CNN	-	-	-	-	68.1	88.1	94.6	-
	DGD	75.3	-	-	-	-	-	-	-
	JLML	83.2	98.0	99.4	99.8	80.6	96.9	98.7	99.2

3.1 Comparisons to State-Of-The-Arts

(I) Evaluation on CUHK03. Table 5 shows the comparisons of JLML against 8 existing methods on CUHK03. It is evident that JLML outperforms existing methods in all categories on both labelled and detected bounding boxes, surpassing the 2nd best performers DGD and X-Corr on corresponding labelled and detected images in Rank-1 by 7.9%(83.2-75.3) and 8.6%(80.6-72.0) respectively. X-Corr/GOG/JLML also suffer the least from auto-detection misalignment, indicating the robustness of the joint learning approach to mining complementary local and global discriminative features.

Table 6: Market-1501 evaluation. $1^{st}/2^{nd}$ best in red/blue. All person bounding box images were auto-detected.

Cat	Query Type	Single-Query		Multi-Query	
	Measure (%)	R1	mAP	R1	mAP
A	XQDA	43.8	22.2	54.1	28.4
	SCS	51.9	26.3	-	-
	NFST	61.0	35.6	71.5	46.0
C	S-CNN	65.8	39.5	76.0	48.4
	JLML	85.1	65.5	89.7	74.5

(II) Evaluation on Market-1501. We evaluated the JLML against four existing models on Market-1501. Table 6 shows the clear performance superiority of JLML over all state-of-the-arts with more significant Rank-1 advantages over other methods compared to CUHK03, giving 19.3%(85.1-65.8) (SQ) and 13.7%(89.7-76.0) (MQ) gains over the 2^{nd} best S-CNN. This further validates the advantages of our joint learning of multi-loss classification for optimising re-id especially when the re-id test population size increases (750 people on Market-1501 vs. 100 people on CUHK03).

Table 7: VIPeR evaluation. $1^{st}/2^{nd}$ best in red/blue.

Cat	Rank (%)	R1	R5	R10	R20
A	XQDA	40.0	68.1	80.5	91.1
	GOG	49.7	-	88.7	94.5
	NFST	51.1	82.1	90.5	95.9
	SCS	53.5	82.6	91.5	96.7
B	DCNN+	34.8	63.6	75.6	84.5
	MTDnet	47.5	73.1	82.6	-
C	MCP	47.8	74.7	84.8	91.1
	DGD	38.6	-	-	-
	JLML	50.2	74.2	84.3	91.6

(III) Evaluation on VIPeR. We evaluated the performance of JLML against 8 strong competitors on VIPeR, a more challenging test scenario with fewer training classes (316 people) and lower image resolution. On this dataset, the best performers are hand-crafted feature methods (SCS and NFST) rather than deep models. This is in contrast to the tests on CUHK03 and Market-1501. Nevertheless, the JLML model remains the best among all deep methods with or without deep verification metric learning. This validates the superiority and robustness of our deep joint global and local representation learning of multi-loss classification given sparse training data. We attribute this property to the JLML's capability of mining complementary features in different context for both handling local misalignment and optimising global matching.

Table 8: GRID evaluation. $1^{st}/2^{nd}$ best in red/blue.

Cat	Rank (%)	R1	R5	R10	R20
A	XQDA	16.6	33.8	41.8	52.4
	GOG	24.7	47.0	58.4	69.0
	SCS	24.2	44.6	54.1	65.2
B	X-Corr	19.2	38.4	53.6	66.4
C	**JLML**	37.5	61.4	69.4	77.4

(IV) Evaluation on GRID. We compared JLML against 4 competing methods on GRID. In addition to poor image resolution, poor lighting and a small training size (125 people), GRID also has extra distractors in the testing population therefore presenting a very challenging but realistic re-id scenario. Table 8 shows a significant superiority of JLML over existing state-of-the-arts, with Rank-1 12.8%(37.5-24.7) better than the 2^{nd} best method GOG, a 51.8% relative improvement. This demonstrates the unique and practically desirable advantage of JLML in handling more realistically challenging open-world re-id matching where large numbers of distractors are usually present.

3.2 Further Analysis and Discussions

We further examined the component effects of our JLML model on the Market-1501 dataset in the following aspects.

Table 9: Complementary benefits of global and local features.

Query Type	Single-Query		Multi-Query	
Measure (%)	R1	mAP	R1	mAP
JLML (Global)	77.4	56.0	85.0	66.0
JLML (Local)	78.9	57.8	86.4	68.4
JLML (joint)	**85.1**	**65.5**	**89.7**	**74.5**

(I) Complementary of Global and Local Features. We evaluated the complementary effects of our jointly learned local and global features by comparing their individual re-id performance against that of the joint features. Table 9 shows that: **(i)** Any of the two feature representations *alone* is competitive for re-id, e.g. the local JLML feature surpasses S-CNN (Table 6) by Rank-1 13.1%(78.9-65.8) (SQ) and 10.4%(86.4-76.0) (MQ); and by mAP 18.3%(57.8-39.5) (SQ) and 20.0%(68.4-48.4) (MQ). **(ii)** A further performance gain is obtained from the joint feature representation, yielding further 6.2%(85.1-78.9) (SQ) and 3.3%(89.7-86.4) (MQ) in Rank-1 increase, and 7.7%(65.5-57.8) (SQ) and 6.1%(74.5-68.4) (MQ) in mAP boost. These results show the complementary advantages of jointly learning the local and global features in different context using the JLML model.

(II) Importance of Branch Independence. We evaluated the importance of branch independence by comparing our *MultiLoss* design with a *UniLoss* design that merges two

Table 10: Importance of branch independence.

Loss	Query Type	Single-Query		Multi-Query	
	Measure (%)	R1	mAP	R1	mAP
UniLoss	Global Feature	58.3	31.7	70.4	43.2
	Local Feature	46.3	26.3	58.0	34.0
	Full	76.1	52.2	83.7	62.8
MultiLoss	Global Feature	77.4	56.0	85.0	66.0
	Local Feature	78.9	57.8	86.4	68.4
	Full	**85.1**	**65.5**	**89.7**	**74.5**

branches into a single loss [Cheng *et al.*, 2016]. Table 10 shows that the proposed MultiLoss model significantly improves the discriminative power of global and local re-id features, e.g. with Rank-1 increase of 9.0%(85.1-76.1) (SQ) and 6.0%(89.7-83.7) (MQ); and mAP improvement of 13.3%(65.5-52.2) (SQ) and 11.7%(74.5-62.8) (MQ). This shows that branch independence plays a critical role in joint learning of multi-loss classification for effective feature optimisation. One plausible reason is due to the negative effect of a single loss imposed on the learning behaviour of both branches, caused by the potential divergence in discriminative features in different context (local and global). This is shown by the significant performance degradation of both global and local features when the UniLoss model is imposed.

Table 11: Benefits from shared low-level features.

Query Type	Single-Query		Multi-Query	
Measure (%)	R1	mAP	R1	mAP
Without Shared Feature	83.2	63.1	88.3	72.1
With Shared Feature	**85.1**	**65.5**	**89.7**	**74.5**

(III) Benefits from Shared Low-Level Features. We evaluated the effects of interaction between global and local branches introduced by the shared conv layer (common ground) by deliberately removing it and then comparing the re-id performance. Table 11 shows the benefits from jointly learning low-level features in the common conv layers, e.g. improving Rank-1 by 1.9%(85.1-83.2) / 1.4%(89.7-88.3) and mAP by 2.4%(65.5-63.1) / 2.4%(74.5-72.1) for single-/multi-query re-id. This confirms a similar finding as in multi-task learning study [Argyriou *et al.*, 2007].

Table 12: Effects of selective feature learning (SFL).

Query Type	Single-Query		Multi-Query	
Measure (%)	R1	mAP	R1	mAP
Without SFL	83.4	63.8	88.7	72.9
With SFL	**85.1**	**65.5**	**89.7**	**74.5**

(IV) Effects of Selective Feature Learning. We evaluated the contribution of our structured sparsity based Selective Feature Learning (SFL) (Eqs. (1) and (2)). Table 12 shows that our SFL mechanism can bring additional re-id matching benefits, e.g. improving Rank-1 rate by 1.7%(85.1-83.4) (SQ) and 1.0%(89.7-88.7) (MQ); and mAP by 1.7%(65.5-63.8) (SQ) and 1.6%(74.5-72.9) (MQ).

Table 13: Comparisons of model size and complexity. FLOPs: the number of FLoating-point OPerations; PN: Parameter Number.

Model	FLOPs	PN (million)	Depth	Stream #
AlexNet	$\mathbf{7.25\times10^8}$	58.3	7	1
VGG16	1.55×10^{10}	134.2	16	1
ResNet50	3.80×10^9	23.5	**50**	1
GoogLeNet	1.57×10^9	**6.0**	22	1
JLML-ResNet39	1.54×10^9	7.2	39	**5**

(V) Comparisons of Model Size and Complexity. We compared the proposed JLML-ResNet39 model with four seminal classification CNN architectures (Alexnet [Krizhevsky *et al.*, 2012], VGG16 [Simonyan and Zisserman, 2015], GoogLeNet [Szegedy *et al.*, 2015], and ResNet50 [He *et al.*, 2016]) in model size and complexity. Table 13 shows that the JLML has both the 2^{nd} smallest model size (7.2 million parameters) and the 2^{nd} smallest FLOPs (1.54×10^9), although containing more streams (5 vs. 1 in all other CNNs) and more layers (39, more than all except ResNet50).

4 Conclusion

We presented a novel Joint Learning of Multi-Loss (JLML) CNN model (JLML-ResNet39) for person re-identification feature learning. In contrast to existing re-id approaches that employ either global or local appearance features alone, the proposed model is capable of extracting and exploiting both and maximising their correlated complementary effects by learning discriminative feature representations in different context subject to multi-loss classification objective functions in a unified framework. This is made possible by the proposed JLML-ResNet39 architecture design. Moreover, we introduce a structured sparsity based feature selective learning mechanism to further improve joint feature learning. Extensive comparative evaluations on four re-id benchmark datasets were conducted to validate the advantages of the proposed JLML model over a wide range of the state-of-the-art methods on both manually labelled and more challenging auto-detected person images. We also provided component evaluations and analysis of the model performance in order to give insights on the JLML model design.

Acknowledgements

This work was partially supported by the China Scholarship Council, Vision Semantics Ltd and Royal Society Newton Advanced Fellowship Programme (NA150459).

References

[Ahmed *et al.*, 2015] Ejaz Ahmed, Michael Jones, and Tim K Marks. An improved deep learning architecture for person re-identification. In *CVPR*, 2015.

[Argyriou *et al.*, 2007] Andreas Argyriou, Theodoros Evgeniou, and Massimiliano Pontil. Multi-task feature learning. In *NIPS*, 2007.

[Chen *et al.*, 2016] Dapeng Chen, Zejian Yuan, Badong Chen, and Nanning Zheng. Similarity learning with spatial constraints for person re-identification. In *CVPR*, 2016.

[Chen *et al.*, 2017a] Weihua Chen, Xiaotang Chen, Jianguo Zhang, and Kaiqi Huang. A multi-task deep network for person re-identification. In *AAAI*, 2017.

[Chen *et al.*, 2017b] Ying-Cong Chen, Xiatian Zhu, Wei-Shi Zheng, and Jian-Huang Lai. Person re-identification by camera correlation aware feature augmentation. *TPAMI*, 2017.

[Cheng *et al.*, 2016] De Cheng, Yihong Gong, Sanping Zhou, Jinjun Wang, and Nanning Zheng. Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In *CVPR*, 2016.

[Edelman, 1998] Shimon Edelman. Representation is representation of similarities. *Behavioral and Brain Sciences*, 21(04):449–467, 1998.

[Farenzena *et al.*, 2010] Michela Farenzena, Loris Bazzani, Alessandro Perina, Vittorio Murino, and Marco Cristani. Person re-identification by symmetry-driven accumulation of local features. In *CVPR*, 2010.

[Girshick *et al.*, 2014] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014.

[Gong *et al.*, 2014] Shaogang Gong, Marco Cristani, Shuicheng Yan, and Chen Change Loy. *Person re-identification*. Springer, January 2014.

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

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[Jia *et al.*, 2014] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. In *ACM MM*, 2014.

[Koestinger *et al.*, 2012] Martin Koestinger, Martin Hirzer, Paul Wohlhart, Peter M. Roth, and Horst Bischof. Large scale metric learning from equivalence constraints. In *CVPR*, 2012.

[Kong *et al.*, 2014] Deguang Kong, Ryohei Fujimaki, Ji Liu, Feiping Nie, and Chris Ding. Exclusive feature learning on arbitrary structures via l1,2-norm. In *NIPS*, 2014.

[Krizhevsky *et al.*, 2012] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

[Kviatkovsky *et al.*, 2013] Igor Kviatkovsky, Amit Adam, and Ehud Rivlin. Color invariants for person reidentification. *TPAMI*, 35(7):1622–1634, 2013.

[LeCun *et al.*, 1998] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *IEEE*, 86(11):2278–2324, 1998.

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

[Liao *et al.*, 2015] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z Li. Person re-identification by local maximal occurrence representation and metric learning. In *CVPR*, 2015.

[Loy *et al.*, 2009] Chen Change Loy, Tao Xiang, and Shaogang Gong. Multi-camera activity correlation analysis. In *CVPR*, 2009.

[Ma *et al.*, 2017] Xiaolong Ma, Xiatian Zhu, Shaogang Gong, Xudong Xie, Jianming Hu, Kin-Man Lam, and Yisheng Zhong. Person re-identification by unsupervised video matching. *Pattern Recognition*, 65:197–210, 2017.

[Matsukawa *et al.*, 2016a] Tetsu Matsukawa, Takahiro Okabe, Einoshin Suzuki, and Yoichi Sato. Hierarchical gaussian descriptor for person re-identification. In *CVPR*, 2016.

[Matsukawa *et al.*, 2016b] Tetsu Matsukawa, Takahiro Okabe, Einoshin Suzuki, and Yoichi Sato. Hierarchical gaussian descriptor for person re-identification. In *CVPR*, 2016.

[Navon, 1977] David Navon. Forest before trees: The precedence of global features in visual perception. *Cognitive Psychology*, 9(3):353–383, 1977.

[Paisitkriangkrai *et al.*, 2015] Sakrapee Paisitkriangkrai, Chunhua Shen, and Anton van den Hengel. Learning to rank in person re-identification with metric ensembles. In *CVPR*, 2015.

[Simonyan and Zisserman, 2015] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

[Subramaniam *et al.*, 2016] Arulkumar Subramaniam, Moitreya Chatterjee, and Anurag Mittal. Deep neural networks with inexact matching for person re-identification. In *NIPS*, 2016.

[Szegedy *et al.*, 2015] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *CVPR*, 2015.

[Torralba *et al.*, 2006] Antonio Torralba, Aude Oliva, Monica S Castelhano, and John M Henderson. Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. *Psychological Review*, 113(4):766, 2006.

[Varior *et al.*, 2016] Rahul Rama Varior, Mrinal Haloi, and Gang Wang. Gated siamese convolutional neural network architecture for human re-identification. In *ECCV*, 2016.

[Wang *et al.*, 2013] Hua Wang, Feiping Nie, and Heng Huang. Multi-view clustering and feature learning via structured sparsity. In *ICML*, 2013.

[Wang *et al.*, 2014a] H. Wang, S. Gong, and T. Xiang. Unsupervised learning of generative topic saliency for person re-identification. In *British Machine Vision Conference*, Nottingham, UK, September 2014.

[Wang *et al.*, 2014b] Taiqing Wang, Shaogang Gong, Xiatian Zhu, and Shengjin Wang. Person re-identification by video ranking. In *ECCV*, 2014.

[Wang *et al.*, 2016a] Faqiang Wang, Wangmeng Zuo, Liang Lin, David Zhang, and Lei Zhang. Joint learning of single-image and cross-image representations for person re-identification. In *CVPR*, 2016.

[Wang *et al.*, 2016b] Hanxiao Wang, Shaogang Gong, Xiatian Zhu, and Tao Xiang. Human-in-the-loop person re-identification. In *ECCV*, 2016.

[Wang *et al.*, 2016c] Taiqing Wang, Shaogang Gong, Xiatian Zhu, and Shengjin Wang. Person re-identification by discriminative selection in video ranking. *TPAMI*, 38(12):2501–2514, 2016.

[Xiao *et al.*, 2016] Tong Xiao, Hongsheng Li, Wanli Ouyang, and Xiaogang Wang. Learning deep feature representations with domain guided dropout for person re-identification. In *CVPR*, 2016.

[Xiong *et al.*, 2014] Fei Xiong, Mengran Gou, Octavia Camps, and Mario Sznaier. Person re-identification using kernel-based metric learning methods. In *ECCV*. 2014.

[Zhang *et al.*, 2016] Li Zhang, Tao Xiang, and Shaogang Gong. Learning a discriminative null space for person re-identification. In *CVPR*, 2016.

[Zhao *et al.*, 2013] Rui Zhao, Wanli Ouyang, and Xiaogang Wang. Unsupervised salience learning for person re-identification. In *CVPR*, 2013.

[Zheng *et al.*, 2013] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Reidentification by relative distance comparison. *TPAMI*, 35(3):653–668, March 2013.

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