Cross-view Discriminative Feature Learning for Person Re-Identification

Alessandro Borgia, Yang Hua, Elyor Kodirov, Neil M. Robertson

***Abstract*—The viewpoint variability across a network of non-overlapping cameras is a challenging problem affecting person re-identification performance. In this paper, we investigate how to mitigate the cross-view ambiguity by learning highly discriminative deep features under the supervision of a novel loss function. The proposed objective is made up of two terms, the Steering Meta Center (SMC) term and the Enhancing Centres Dispersion (ECD) term that steer the training process to mining effective intra-class and inter-class relationships in the feature domain of the identities. The effect of our loss supervision is to generate a more expanded feature space of compact classes where the overall level of inter-identities interference is reduced. Compared to the existing metric learning techniques, this approach has the advantage of achieving a better optimization because it jointly learns the embedding and the metric contextually. Our technique, by dismissing side-sources of performance gain, proves to enhance the CNN invariance to viewpoint without incurring increased training complexity (like in Siamese or Triplet networks) and outperforms many related state-of-the-art techniques on Market-1501 and CUHK03.**

***Index Terms*—Viewpoint, Loss function, Multi-camera, Person re-id, Discriminative features**

I. Introduction

PERFORMING person re-identification is a challenging task because of the presence of many sources of appearance variability like lighting, pose, viewpoint, occlusions, especially in outdoor environment [1], [2] where they are even more unrestrained. Cameras calibration or cross-camera image processing may help, but they are not an option in a surveillance context where a *wide-area network of cameras with non-overlapping fields of view* is deployed. In such a scenario, we investigate the *changing viewpoint problem*. The misleading effect of this particular factor of variability is that shots of different pedestrians taken under the same camera may quite often look more similar to each other than shots of the same identity taken under different cameras. This is illustrated in Figure 1. We support the view that learning inter-camera relationships is essential to tackle this ambiguity [3], [4] since it can contribute to build, at training time, a more discriminative feature space where all classes (pedestrian identities) are less conflated and more distant from each other (Figure 2). It can be done by properly designing a loss function, tailored to the goal, that supervises the learning process of a deep architecture.

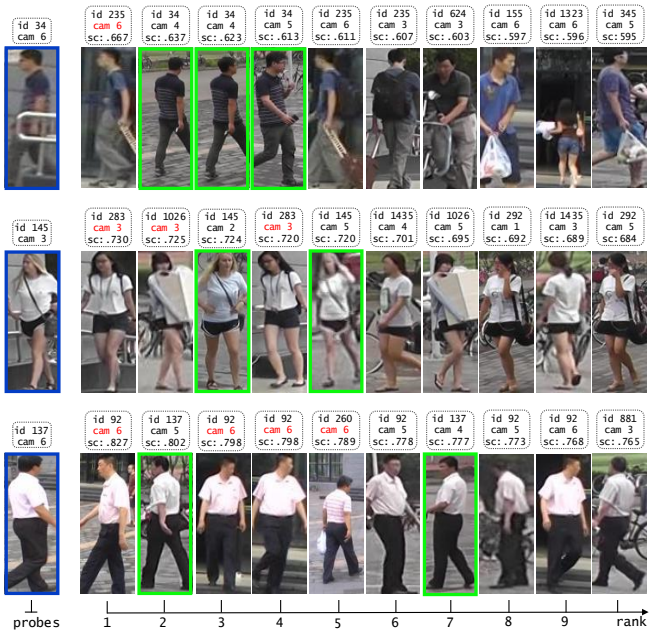


Fig. 1. Viewpoint problem in re-id (Market-1501 dataset). For each probe (blue framed), the list of the 10 top-ranked images is reported (green framed images successfully match the probe identity). All the negatives with a red camera label (referred to as "negatives of interest") express an occurrence of the viewpoint problem, meaning that they are false positives shot by the same camera of the probe. All figures are best viewed in color.

Apart from some early techniques relying on designing hand-crafted features [5] or cross-camera transformations [3], the multi-camera context, traditionally, has embraced the deep learning (DL) paradigm because of its ability to learn complex discriminative mappings that generalise well [6]–[10]. Most of the DL approaches in person re-id focus on exploiting one of the following aspects: **1)** more complex deep architecture structures, aiming either to optimize jointly more tasks in the re-id pipeline [11], [12] or to learn cross-spatial/temporal representations [13]; **2)** side information extraction, [14], [15], involving pose estimation strategies and misalignment correction; **3)** metric learning [16], [17] that learns a distance/similarity function in a fixed feature space; **4)** more training data, either by performing cross-dataset training [18] or exploiting the transfer learning paradigm [19].

With regards to the above classification, our approach based on employing a more discriminative loss function, has the ad-

A. Borgia, corresponding author, is with ISSS, Heriot-Watt University and SIP-JRI, University of Edinburgh, UK. Email: ab41@hw.ac.uk

Y. Hua, corresponding author, is with EEECS/ECIT, Queen's University Belfast, UK. Email: y.hua@qub.ac.uk

E. Kodirov is with Anyvision, Belfast, UK. Email: elyor@anyvision.co

N. M. Robertson is with EEECS/ECIT, Queen's University Belfast, UK. Email: n.robertson@qub.ac.uk

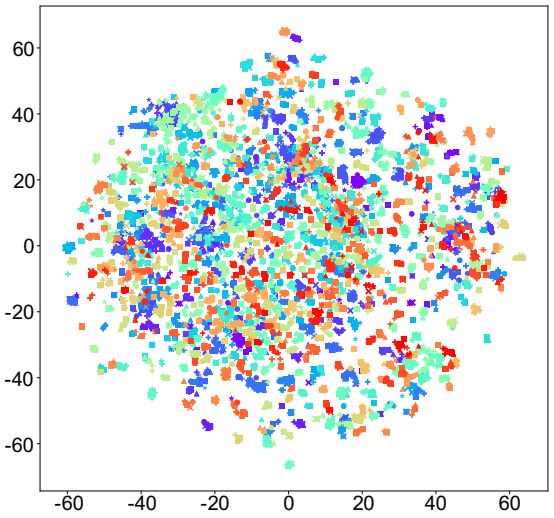


Fig. 2. 2D visualization of the entire Market-1501 test set feature space when only the softmax supervision is applied (T-Sne visualization tool [22]). Each point is the 2D projection of the multi-dimensional feature of one image of the dataset. Colors represent identities, while markers denote camera views. The axes scale measures the size (by normalized Euclidean distance) of the dataset representation. Features points of resembling images are gathered in overlapped clusters.

vantage of being totally complementary to the these methods: it can easily be integrated into any architecture that leverages a higher structural complexity, enhanced data exploitation or ML. Substantially, our work investigates how to make the most out of a fixed available amount of training data, without relying on the exploitation of any side information [20], [21]. A notable example of this technique in face verification is [20] that proposes the *center loss* function that enhances the softmax loss supervision by promoting the compactness of the extracted features around the center point of each face-class. Although the idea is quite effective in face verification, we think that its limitation in person re-id is that it does not exploit the field of view information, available in multi-camera datasets, which could be helpful to mitigate the viewpoint problem. [21] tries to fill this gap with the intra- and inter-Group-Center Losses (GCL): it incorporates the field of view information in their definition, although they still face some limitations because of their mathematical formulation (Sec.III). In our work we overcome these limitations by proposing a novel loss made up of two terms, the Steering Meta-Center term and the Enhancing Classes Dispersion term. Its goal is to exploit effectively the field of view information in order to foster the separation capability of the softmax loss and enhance further the intra-class gathering behaviour of the center loss.

In summary, the main contributions of our work are the following:

- We propose a flexible method to successfully mitigate the viewpoint problem in person re-id by tailoring a new loss function that enables a CNN to learn more discriminative features.
- Our approach adapts the ML approach to the training stage for learning an inter-class distance/similarity function.
- Our results beat the best performing loss function-based approaches (Table III) and most of the state-of-the-art methods (Table II) on two of the largest datasets for outdoor person re-id.

II. Related Work

A. Deep learning for person re-id

The first work to apply DL to person re-id is [6] that relies on a Siamese CNN equipped with a cosine similarity connection function. Following this research direction, lots of other works, taking advantage of the availability of new large-scale datasets (CUHK03 [7], Market-1501 [23]), have adopted either the pure data-driven paradigm [8], [10], [18] or hybrid approaches [7], [24]–[26]. In mixed architectures, hand-crafted descriptors for pedestrians are integrated into DL frameworks and exploit the complementarity of their features. Fisher vectors and deep neural networks are combined together, for example, in [24]; [7], [25] design a Siamese network with constraints on the shape of the objective to learn by adding custom hand-crafted layers to the CNN; in [26] convolutional and hand-crafted histogram features are fused together to produce a more discriminative descriptor. In order to enhance performance in deep architectures for person re-id, two popular strategies exploit either some spatial cues of the input images [27] or the side information extracted from data, like in [14] where, within a DL framework, an effective correction of full-body images misalignment is performed by using Convolutional Pose Machines for pose estimation. A recent trend consists in addressing different re-id related tasks jointly. In this perspective, [12] in its multi-task network fuses the binary classification and ranking tasks together, while [11] integrates the pedestrian detection and searching tasks in one unified end-to-end trainable net, taking a significant step ahead in the direction of filling the gap between research-oriented re-id systems and real world deployable re-id systems. Lastly, transfer learning-based architectures provide evidence of how much performance benefits from the availability of extra data in person re-id by learning generic deep feature representations from multiple domains [18], [19]. In our work we adopt the ResNet50 architecture also used in [14] which relies on the effectiveness of the residual learning paradigm for learning deeper feature hierarchies [28].

B. Loss functions

The simple probabilistic interpretation of the softmax loss and its features separation capability make it the most widely used loss [26], [29]. In multi-label prediction, the sigmoid cross-entropy classification loss is sometimes preferred for its better performance like in [30] where it is used for jointly learning correlated complementary local and global features. Common approaches in this area involve either modifying the softmax loss or replacing/combining it with novel losses. The shared goal is to enforce intra-class compactness and inter-class separability in the feature space.

A modification of the softmax loss is proposed in [11] with the random sampling softmax loss that allows supervising the

training stage with sparse and unbalanced labels. In [31] the generalized large-margin softmax loss generalizes the softmax loss incorporating in its new definition the cross-entropy loss and a fully connected layer to achieve larger angular feature separability. The reduction of intra-class variability has been addressed in [20] by the *center loss* that learns the centers of deep features of each id-class (in face verification), and in [21] by the two combined *GCL* losses that also promote inter-class separation.

The choice of which loss function to use may heavily affect the way the training samples are built, having an impact eventually on the training complexity. This is the case for the contrastive loss [32] which is used in the Siamese network model. It proves to be quite effective in person re-id like in [8] and [10] that consider ways of exploiting spatial relations of images, within a single image or between image pairs. A boosted learning capability derives from including into a Siamese framework more losses related to different visual tasks (identification and verification) combined together [19].

A direct competitor of the Siamese loss is the triplet loss for the triplet network model [33], [34] that enables insensitivity to calibration which is a problem in Siamese CNNs where the concept of similarity/dissimilarity is tied to the specific context [35]. In [9], an improved triplet objective is used with an upper-threshold on the maximum distance for the intra-class features, in order to train a multi-channel parts-based CNN. A combination of the pairwise and triplet-wise feature learning modality is also presented in [13], [36].

C. Learning a metric in the feature space

Distance-ML based methods learn transformations of the original feature space in order to project embedded representations belonging to different cameras onto a common space where the view discrepancy is mitigated [37], [38]. Typically, the existing ML approaches may be classified in two groups: non-DL ML methods and DL-based ML methods.

The first group is characterized by performing feature embedding learning and ML in two separated subsequent stages, so that a metric is learned only after the CNN weights are already fixed. In this group falls [39] that applies deep transfer ML to learn a set of hierarchical non-linear transformations for cross-domain recognition. Intra-person variability in [40] is handled by carrying out a similarity learning that obeys some spatial constraints accounting for the geometrical structure of pedestrians (to match correspondent body-parts). [41] learns a new metric as a combination of a Mahalanobis metric and a bilinear similarity metric and benefits from considering both the difference and commonness of an image pair and a pair-constrained Gaussian assumption. An original definition of similarity metric is presented in [42] as a log likelihood ratio between the probabilities of the two identities to be matched (in face recognition).

In contrast, in the second group of ML methods, a metric is learned jointly with the embedding by a DL-based architecture. One way to achieve this is in a Siamese network fashion, exploiting the contrastive loss (or triplet loss) that allows a distance relation between pairs (or triples) of feature points to be learned at training time [6], [9], [43], [44]. Another approach among the DL-based ML methods is to integrate the ML scheme into the feature extraction CNN, producing a re-id system overall trainable end-to-end by gradient descent. In [16], [38], for example, the Mahalanobis distance matrix is factorized as one top fully-connected layer and the matrix of its weights is learned together with the CNN weights. Similarly, [45] finds an end-to-end globally optimal matching in a multi-camera network by exploiting intra- and inter-camera consistent-aware information during the training stage of a three branches CNN.

Both the Siamese-based and the end-to-end integration-based approaches often need to be supported by *sample mining strategies* to either improve the training effectiveness [8], [25] or reduce the over-fitting problem [16]. A hard negatives mining strategy is applied for example in [25] and in [7] where, after being retrieved, the hard negatives are combined with the positive samples to further train the network. Moderate positive samples selection is required instead in [16] to mitigate the increased over-fitting to bad positives. The loss we propose allows us to avoid dealing with any sample mining strategy while, at the same time, benefiting from the embedding-metric joint learning.

III. Proposed Method

Aiming to reduce the viewpoint ambiguity connected to the multi-camera scenario by producing more discriminative deep features, we assert that learning a metric only over a fixed CNN is limiting compared to conveniently shaping the feature space itself while it gets built, at training time. Therefore an approach is required that address the two tasks of *learning an embedding* and of *learning a metric* jointly, avoiding the main drawbacks of the Siamese architectures, but still retaining their inter-camera relationships learning capability, critical for enabling the viewpoint invariance.

When ML is performed separately by the feature extraction task, besides reaching a suboptimal solution, usually, either dimensionality reduction or regularization are required to avoid singularity in the intra-class scatter matrix due to the limited number of training samples for a single identity compared to their feature dimensionality (*small sample size* problem, [46]). On the other hand, the joint learning performed by a Siamese network-based approach may result in an increased training complexity for several reasons. Firstly, the explosion of the number of samples due to the need to build the training samples by selecting pair/triplet of input images. Secondly, the need of sample mining strategies to improve the effectiveness of the training [8], [25]. Thirdly, the contrastive loss only relies on weak re-id labels (same id or different id), as pointed out in [47], and does not exploit fully the entire information carried by the re-ID labels on class membership. Lastly, because of the *unbalanced training data* problem, by learning a network by binary classification, their predictions are usually biased towards negatives. Countermeasures to this, usually require to increase the training complexity even further as in [48] where a unified deep learning-to-rank framework is proposed.

Our approach replicates the capability of Siamese networks to carry out a joint features-metric learning process while

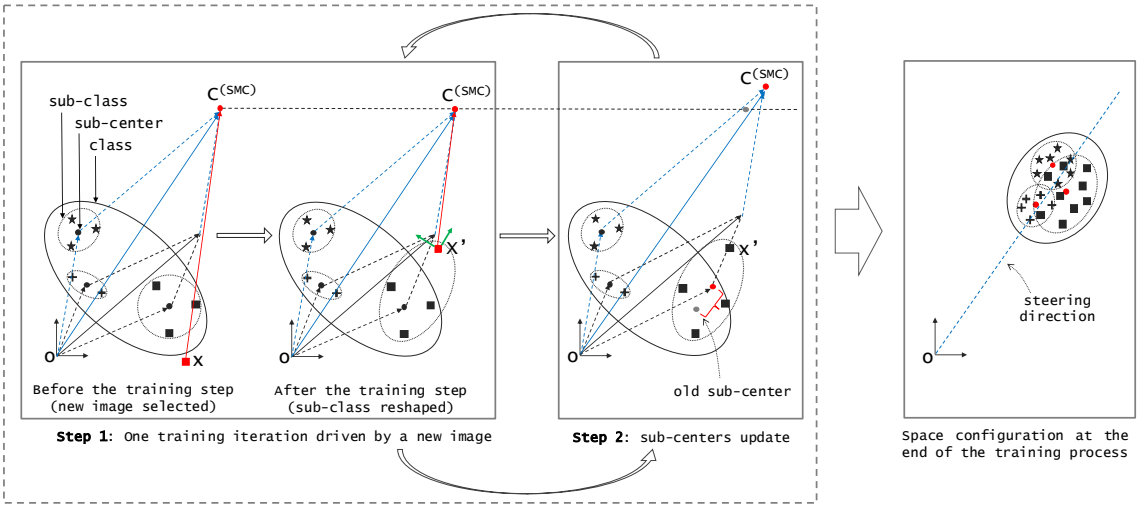


Fig. 3. Representation of our approach to the iteration-based optimization of the SMC local loss and its overall effect on the feature space. One class represents an identity; its sub-classes represent the camera views; the centers of the id sub-classes are the class sub-centers; $\boldsymbol{x}$ denotes a new data point and $\boldsymbol{x'}$ its novel position in the updated space; $C^{(SMC)}$ is the meta-center. The potential attracting $\boldsymbol{x}$ to $C^{(SMC)}$ (red arrow) may be seen as the resultant of its two components (green arrows): one potential *drifting* $\boldsymbol{x}$ away from the origin and one potential *compressing* the points cloud of the considered identity towards the steering direction identified by $C^{(SMC)}$. The system origin is assumed to be the center of the dataset points cloud.

at the same time keeps the training complexity low, since one training sample corresponds to a single input image, like in the non-DL metric learning methods. We design a novel loss function that has the nice properties of a) being additive with regards to the softmax loss; b) being suitable to be easily integrated in a simple one branch shaped CNN, being trainable by gradient descent; c) being suitable for fast search requirements since it scales well to large datasets; d) producing embeddings discriminative enough that simple metrics such as the normalized Euclidean distance can be used for comparing the multi-dimensional feature points representing the identities instances.

The loss function we design is made up of two additive terms: the Steering Meta-Center term (SMC) and the Enhancing Classes Dispersion term (ECD). Used in linear combination with the softmax loss, they promote two desirable properties of the deep features distribution: the properties of intra-class compactness and of inter-class separation. In particular we test them in two combinations: SMC+ECD as in Equation 1 and SMC only, without the ECD term contribution. Beyond producing a more discriminative feature space under the SMC+ECD loss supervision, we investigate also the relationship between our method and the traditional ML approach, in order to understand whether combining them together a further improvement can be gained (Sec. IV-C).

$$L = L_{softmax} + \lambda_{SMC} \cdot L_{SMC} + \lambda_{ECD} \cdot L_{ECD} \qquad (1)$$

A. Steering Meta-Center (SMC) Loss Term

The SMC loss definition exploit the camera information of all the dataset images aiming at two goals: a) Improving the center loss [20] compactness in person re-id; accounting for the camera information helps, indeed, balance the contributions of the different views to defining a more discriminative deep representation of the overall identity. b) Learning to some extent inter-camera relationships, which allows to outdistancing different identities. This task is usually deferred to after the training stage and performed by metric learning schemes.

In the following, we will illustrate how these two aspects are dealt with jointly by the SMC loss term, by introducing a new virtual point in the feature space, referred to as *meta-center* which steers the learning and shapes the feature space. With regards to an identity, a meta-center point is defined simply as the vector sum of its sub-centers, as clear from Equation 2, defining the SMC loss,

$$L_{SMC} = \frac{1}{2}\sum_{i=1}^{m} \|\boldsymbol{x}_i^{(g_i)} - \sum_{j=1}^{s_i} \boldsymbol{c}_{y_i}^{(j)}\|_2^2 = \frac{1}{2}\sum_{i=1}^{m} \|\boldsymbol{x}_i^{(g_i)} - \boldsymbol{c}_{y_i}^{(SMC)}\|_2^2 \qquad (2)$$

where m is the training mini-batch size; y_i denotes the identity ground-truth label of the i^{th} mini-batch image; s_i represents the number of cameras that capture the identity y_i; g_i is the camera ground-truth label of the i^{th} mini-batch image ($1 \leqslant g_i \leqslant s_i$); $\boldsymbol{x}_i^{(g_i)}$ denotes the feature vector of the i^{th} input image viewed under camera g_i ; $\boldsymbol{c}_{y_i}^{(j)}$ represents a sub-center point of the identity y_i, calculated by averaging the points of y_i that are viewed under camera j; $\boldsymbol{c}_{y_i}^{(SMC)}$ is the sum of all the s_i camera-related sub-centers $\boldsymbol{c}_{y_i}^{(j)}$ of identity y_i.

It should be noted that Equation 2 does not represent a global objective to be minimized by the overall training process; instead, its minimization is carried out locally, with scope limited to each individual iteration (similarly to what happens for the triplet loss), with regards to the current value only of the meta-center.

At training time, each new training image being processed is pulled toward the meta-center of its class according to Equation 2. The subsequent step consists in updating the position of the correspondent sub-center that, in turn, moves

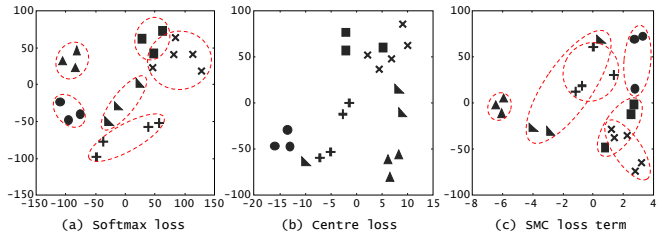


Fig. 4. Effect of the SMC loss term (Market-1501): the increased feature space compactness achieved by SMC vs softmax loss and center loss is visible from the decreasing axes range in the three different 2-D T-Sne projections (a),(b) and (c) of the same identity. The dashed circles highlight clusters of feature points corresponding to the same camera view. In (c) the sub-clusters appear more overlapped than in (a).

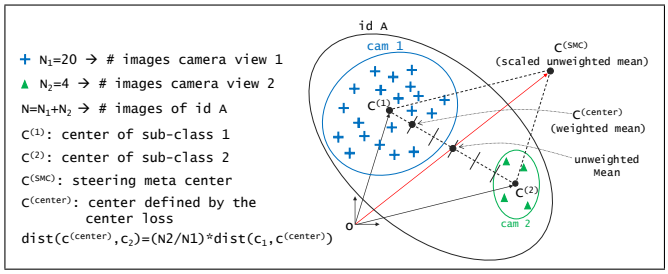


Fig. 5. Relation between the center defined by the center loss and by the SMC loss. The two sub-classes cam 1 and cam 2 of identity A are unbalanced since $N_1 >> N_2$. The center C_c defined by the center loss corresponds to the weighted mean (by N_1 and N_2) of the sub-centers and is very near to the sub-center of the prevalent camera view. In contrast, the unweighted center treats equally both cam 1 and cam 2 addressing the sub-class invariance. The meta-center retains the sub-class invariance property of the unweighted center representing simply its scaled version. Considering a scaled version of the unweighted center (for a factor $= N > 1$) that falls out of the identity class is convenient in that it adds the inter-class dispersion behaviour (according to Figure 3) besides addressing intra-class compactness.

away from the reference system origin, following the meta-center (Figure 3). The repetition of this training cycle, on the one hand, makes the feature points of all identities drift away from the system origin; on the other hand, it makes each class tend to approach more and more tightly the steering direction of the meta-center. The overall effect of these two potentials is to simultaneously:

- produce a more scattered space of classes because of the identities progressive movement behind their own meta-centers.
- achieve a high intra-class compactness and a less sub-clustered space structure (more inter-mixed points regardless of their camera view), as a result of the increased insensitivity to the camera viewpoint (Figure 4).

The training process is concurrently supervised by the softmax loss and lasts until the produced features start losing generalization power due to data over-fitting.

It is interesting to analyze the mathematical relation between the meta-center point $\boldsymbol{c}^{(SMC)}$ defined by the SMC loss and the center point $\boldsymbol{c}^{(center)}$ defined by the center loss, in order to gain a deeper insight into their difference. In Equation 3 we reformulate the center loss as a function of the sub-centers variables, for a fixed identity,

$$\boldsymbol{c}^{(center)} = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{x}_i = \frac{1}{N}\sum_{j=1}^{s_i}\sum_{i=1}^{N_j}\boldsymbol{x}_i^{(j)} = \frac{1}{N}\sum_{j=1}^{s_i}N_j\boldsymbol{c}^{(j)} \tag{3}$$

where N is the number of images of the identity of interest y_i shot by s_i cameras; N_j is the number of images of the considered identity belonging to the camera view j; $\boldsymbol{c}^{(j)}$ is the sub-center associated to camera view j. This formulation shows that $\boldsymbol{c}^{(center)}$ is equivalent to the *weighted mean* of the sub-centers of the class, with weights given by the cardinality of the population of each sub-class. In contrast, the meta-center, defined in Equation 2 as $\boldsymbol{c}^{(SMC)} = \sum_{j=1}^{s_i}\boldsymbol{c}^{(j)}$ is a scaled version, by a factor N, of the *unweighted mean* of the class sub-centers $\frac{1}{N}\sum_{j=1}^{s_i}\boldsymbol{c}^{(j)}$, as clear from Figure 5. Therefore, the SMC loss has the nice property of referring to the unweighted mean instead of the weighted mean which allows to account

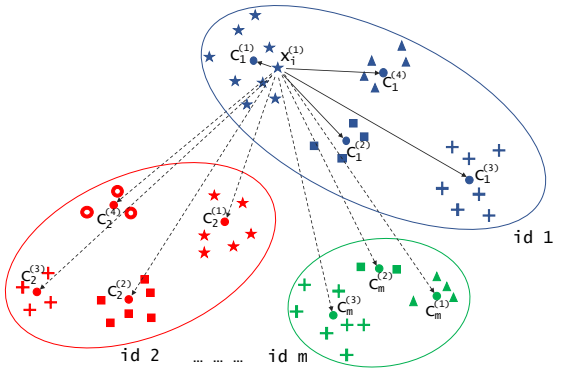


Fig. 6. Representation of how the ECD loss term works. Id 1 is the reference identity, namely the identity that the training point $\boldsymbol{x}_i$ belongs to, in the current mini-batch of size m. The sum of the solid lines distances defines the *class range* of the reference class and is expressed mathematically by the left-hand term of the product in Equation 4. The dotted line distances (*sub-center distances*) account for the proximity of the reference class to the others and are formally expressed by the denominator of the right-hand term of the same product. The ECD loss aims to make small the ratio between the class range and each individual sub-center distance in the mini-batch.

for an equal contribution from all the sub-classes in defining an identity representation, regardless of the number of points they have. Since we want to address the invariance of an identity representation from the camera views, we give all the sub-classes the same relevance in determining the point that summarizes the class.

The problem of accounting for the camera information as a mean to improve the center loss class compactness has already been attempted by the intra-GCL loss [21]. By the way, its formulation shows some limitations due to the fact that it does not constrain the sub-centers of one identity (which condense the camera information) to converge to each other and the effect is to get compact sub-classes in a still wide overall class with consequently reduced performance.

B. Enhancing Classes Dispersion (ECD) Loss Term

The ECD loss is designed to enforce explicitly higher inter-class dispersion by imposing a relative constraint between

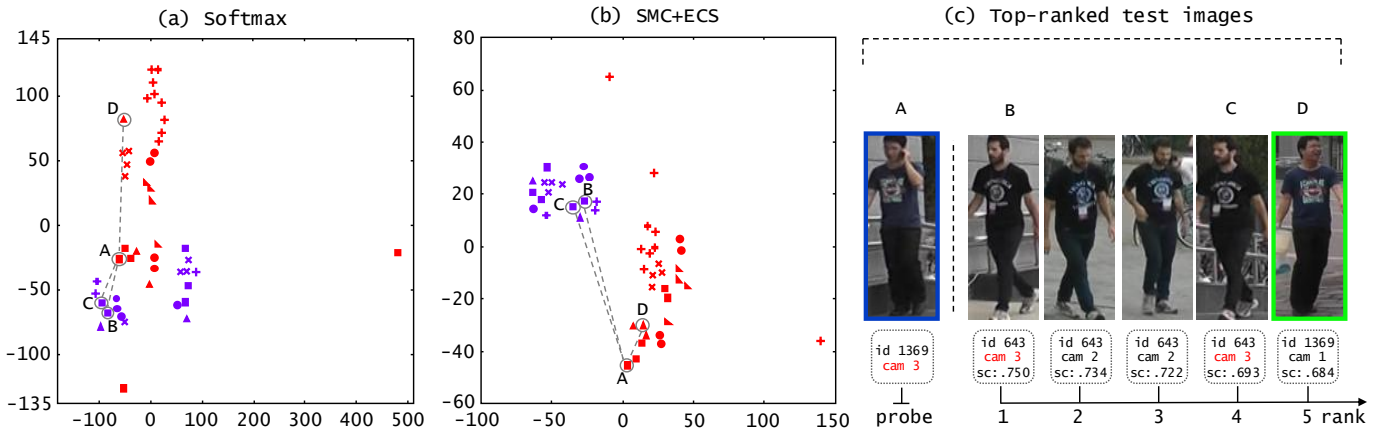


Fig. 7. Effect of the supervision of the SMC+ECD combined loss (Market-1501). With regards to the re-identification of the probe image with id 1369 and camera view 3 (blue framed), in the list of the corresponding top ranked test images (c), the bounding boxes B (rank 1) and C (rank 4) precede the first right match (bounding box D, rank 5, green framed) showing a *viewpoint problem* occurrence. Indeed, images B and C, differently from the correct match D, are viewed under the same camera of the probe (cam 3) and that is the reason why they rank higher. This causes B and C to be nearer to the probe A in the feature space than image D when only the softmax loss is used (a). Performing a training with the SMC+ECD loss reverses the situation (b).

intra-class scope distances and inter-class scope distances. Its formulation is reported in Equation 4: the left-hand term of the product accounts for how much extended the reference class is in the feature space (class range); the right-hand term, instead, expresses a measure of the isolation of the reference identity from each of the other identities in the current mini-batch (Figure 6).

$$L_{ECD} = \frac{1}{2}\sum_{i=1}^{m}\left[\sum_{j=1}^{s_i}\|\boldsymbol{x}_i^{(g_i)} - \boldsymbol{c}_{y_i}^{(j)}\|_2^2 \cdot \sum_{\substack{t=1\\ t\neq i}}^{m}\sum_{k=1}^{s_t}\frac{1}{\|\boldsymbol{x}_i^{(g_i)} - \boldsymbol{c}_{y_t}^{(k)}\|_2^2}\right] \tag{4}$$

Our starting point in the definition of Equation 4, is noting that the inter-GCL loss minimization [21] may not be effective in case the centers initialization values were subject to a large spread with respect to the data (unbalanced centers): namely, $\|\boldsymbol{x}_i^{(g_i)} - \boldsymbol{c}_{y_t}^{(k)}\|_2^2$ very large for some y_t and k. Under this circumstance, indeed, it is interesting to analyze how better the behaviour of the right-hand term of the product of Equation 4 is compared to the counterpart term of the GCL loss, reported in Expression 5,

$$\frac{1}{\sum_{t=1}^{n}\sum_{\substack{j=1\\ j\neq g_i}}^{s_t}\|\boldsymbol{x}_i^{(g_i)} - \boldsymbol{c}_{y_t}^{(j)}\|} \tag{5}$$

where n is the overall number of images in the entire dataset. The inter-GCL expression flatten to zero when even only one sub-center of one class is badly initialized, affecting negatively the function minimization. This behaviour is due to the fact that the inter-GCL term in Equation 5 expresses a weak constraint, forcing the class range to be small only with respect to the sum of the inter-class distances of all the identities in the dataset.

Differently, our solution bypasses this problem by constraining the class range to be small with regards to each individual sub-center distance at a time. In the ECD loss formulation

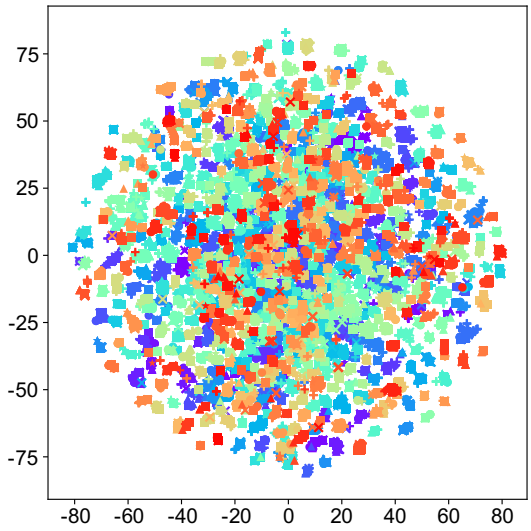


Fig. 8. 2D visualization of the entire Market-1501 test set feature space after the joint supervision of the softmax and the SMC+ECD losses. Respect to the original feature space illustrated in Figure 2, this one appears more expanded, with an increase of about $1/3$ of the original size along each axis. Furthermore, the more crowded appearance of the expanded space is due to the presence of a higher number of better disentangled identities which originally where totally occluded by other similar ones.

(Equation 4), even if several terms of the summation flatten to zero due to some unbalanced centers, the other addends finite contribution will not be affected. Furthermore, another difference is that we limit the summation of the inter-class distances only to the identities in the current training mini-batch instead of accounting at once for all the dataset identities.

The larger the number of the sub-classes is, the stronger the effect of our constraint is expected to be, how confirmed from the best results achieved on Market-1501 dataset than on CUHK03. As to the viewpoint problem, learning a CNN under the joint supervision of the SMC and ECS loss terms produces a positive impact, as illustrated in Figure 7 for a small subset

TABLE I
NETWORK STRUCTURE: RESNET50 SUPERVISED BY THE SMC+ECD LOSS. EACH LAYER AGGREGATE DENOTED BY "CONVX" INCLUDES ONE SPECIFIC RESIDUAL BLOCK REPEATED SEVERAL TIMES.

layer name	output size	kernel size	stride	pad
conv1	112x112	[7x7, 64]	2	3
max pooling	56x56	[3x3]	2	-
conv2	56x56	[1x1, 64; 3x3, 64; 1x1, 256] x3	[1; 1; 1] x3	[0; 1; 0] x3
conv3	28x28	[1x1, 128; 3x3, 128; 1x1, 512] x4	[2; 1; 1] x4	[0; 1; 0] x4
conv4	14x14	[1x1, 256; 3x3, 256; 1x1, 1024] x6	[1; 1; 1] x6	[0; 1; 0] x6
conv5	7x7	[1x1, 512; 3x3, 512; 1x1, 2048] x3	[1; 1; 1] x3	[0; 1; 0] x3
avg pool5 (deep feat)	1x1	[7x7]	1	-
fc8	1260/751	2048	-	-
softmax	1	1260/751	-	-
SMC, ECD	1	2048	-	-

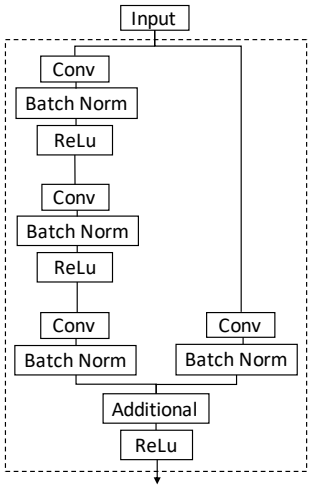


Fig. 9. Residual block structure in ResNet50. The input block represents a mini-batch of feature maps coming either from the max pooling layer or from a previous residual block.

of identities and in Figure 8 for the whole dataset (Market-1501). The ECD loss reproduces at training time what non-DL ML methods do on top of a CNN already learned, learning a distance relation between inter-class pairs.

IV. EXPERIMENT

A. Settings

Database. We evaluate our approach against two of the largest person re-id dataset: **CUHK03** [7] and **Market-1501** [23]. With regards to CUHK03, all results refer to its labeled subset (Table II). In CUHK03, each identity is shot by one pair of cameras out of the three pairs available ($s_i = 2$ in Equation 2) and counts maximum 10 images: the first 5 images are viewed under a different camera than the remaining ones. In Market-1501, each identity is seen under up to 6 different views ($3 \leqslant s_i \leqslant 6$) for up to 70 images. The 12936 images of the train set correspond to 751 identities completely disjoint from the 750 identities of the test set, having 13115 instances. 2798 more images representing heavily misaligned detections with ID identifier '0000' (distractors) are added to the test set in order to make the re-identification task more challenging. The database includes as well a query set of 3368 images (probes) which are picked from the test set so that, for each id, only one image per camera view is selected (mislabeled examples reported in Appendix Table VII). The images of both the datasets are generated by the DPM detector [49].

Evaluation Metric and Protocols. Our ranking-based evaluation method is conducted by matching (by cosine similarity) the feature vectors of all the test images against a probe representation and then sorting the correspondent similarity scores in decreasing order (Figure 7-c). The common evaluation metric we use for measuring performance against both the datasets is the Cumulative Matching Curve (**CMC**, Table III). For Market-1501 we also employ the mean-Average Precision (**mAP**, Table III) since it has multiple ground-truths for each query and both precision and recall need to be taken into account.

Market-1501 comes in the form of three directories corresponding to train set, test set and query set. We follow its original evaluation protocol [23] for which all the query images have to be tested against their own gallery set. Each gallery set excludes the test images that have a filename starting with '-1' and that belong to the probe junk set (made up of all the test images sharing the probe same identity and field of view). Our experiments are performed both in **single query** testing mode (results in Table II and Table III) and in **multiple query** mode (Table IV). As to the former, only one query image is selected for each camera view of a given id. In the multiple query mode, the presence of multiple query images in a single camera for each identity allows to achieve superior performance in re-identification [50]. We apply the max pooling strategy (which performs better than the average pooling) to merge them into a single query for speeding up the process, as done in [23]. For CUHK03, we reproduce the evaluation protocol in [7], according to which the first 5 images of each identity represent view A, the remaining 5 represent view B. View A includes camera 1, 3, 5 while view B camera 2, 4, 6. All images belonging to view A form the probe set. To each probe corresponds a gallery set of 100 images selected randomly from view B, such that one image is picked for each of the 100 identities in the test set. This selection mode of the gallery set is repeated 50 times in order to calculate the mean CMC curve. Each of the 20 test sets counts 100 images and the validation set includes 100 identities.

Besides the rank 1 accuracy and the mAP, we introduce the **figure of merit** $F^{(L)}_{rank1}$ in order to quantify which fraction of the overall improvement achieved by a new loss L translates, specifically, in an improvement of the viewpoint effects. $F^{(L)}_{rank1}$ is associated with the rank 1 accuracy metric (likewise, we can define $F^{(L)}_{mAP}$) and its definition comes

TABLE II
COMPARISON OF OUR METHOD AGAINST SOME OF THE MOST POPULAR TECHNIQUES IN PERSON RE-ID ON MARKET-1501 AND CUHK03.

	Market-1501			CUHK03
Method	**rank1**	**mAP**	**Method**	**rank1**
PersonNet [44]	37.21	18.57	CDM [16]	40.91
DADM [51]	39.40	19.60	Basel.(R, pool5) [14]	51.60
Multiregion CNN [43]	45.58	26.11	SI-CI [13]	52.17
Bow + HS [23]	47.25	21.88	DCNN [25]	54.74
Fisher Network [24]	48.15	29.94	DARI [38]	55.4
SL [40]	51.90	26.35	LSTM Siam. [8]	57.3
DNS [46]	61.02	35.68	PIE(A, FC8) [14]	62.4
LSTM Siam. [8]	61.6	35.3	DeepDiff [52]	62.43
Gated S-CNN [10]	65.88	39.55	DNS [46]	62.55
P2S [36]	70.72	44.27	Fisher Network [24]	63.23
Basel.(R, Pool5) [14]	73.02	47.62	Multiregion CNN [43]	63.87
CADL [45]	73.84	47.11	PersonNet [44]	64.80
PIE(R, Pool5) [14]	78.65	53.87	Gated S-CNN [10]	68.10
ours	**80.31**	59.68	**ours**	**69.55**

TABLE III
CMC SIGNIFICANT POINTS (%) FOR LOSS FUNCTION-BASED METHODS.

	Market-1501 [23]					**CUHK03** [7]			
	mAP	**rank**				**rank**			
		1	**5**	**10**	**20**	**1**	**5**	**10**	**20**
Softmax	47.62	73.02	85.84	90.35	93.32	51.60	79.60	87.70	95.00
GCL	54.25	76.63	88.78	92.25	95.19	63.66	88.58	94.20	98.03
Center	57.76	78.77	90.14	93.62	95.72	66.19	90.65	96.06	98.73
SMC	58.28	79.51	90.59	93.74	95.90	**69.59**	92.62	96.86	98.91
SMC+ECD	**59.68**	**80.31**	91.27	94.09	96.02	69.55	90.96	95.07	97.54

TABLE IV
PERFORMANCE (%) IN THE MULTIPLE QUERIES EVALUATION MODE FOR MARKET-1501 DATASET.

	softmax	**GCL**	**center**	**SMC**	**SMC+ECD**
rank 1	80.94 (+10.8)	81.56 (+6.4)	85.01 (+7.9)	85.14 (+7.1)	85.63 (+6.6)
mAP	57.45 (+20.6)	61.74 (+13.8)	66.01 (+14.4)	66.08 (+13.4)	67.28 (+12.7)

from the analysis of the negatives of interest produced by the ranking process. If we choose the softmax loss as our baseline, $F^{(L)}_{rank1}$ can be defined as: $F^{(L)}_{rank1} = \frac{Neg^{(soft)} - Neg^{(L)}}{rank1^{(L)} - rank1^{(soft)}}$, where $Neg^{(soft)}$ ($\backslash Neg^{(L)}$) is the percentage of *negatives of interest* (Figure 1) produced under the supervision of the softmax ($\backslash$L) loss; $rank1^{(soft)}$ ($\backslash rank1^{(L)}$) represents the performance achieved by the softmax ($\backslash$L) loss (Table V).

Also we build the **mAP camera-pairs confusion matrix** (Figure 10) for the SMC+ECD loss in Market-1501 against the baseline represented by the single softmax loss function, in order to measure the relative improvement of re-id between camera pairs, that is of the viewpoint problem. With regards to a pair of cameras (X, Y), representing respectively the field-of-view of the probe and a test camera, the corresponding mAP value for the considered probe is calculated by limiting its positive samples to only those ones viewed under camera Y. This process is repeated for all probes and values corresponding to the same camera pair are averaged together to produce a single value for the related cell in the matrix. Camera pairs sharing more similar characteristics are characterized by higher values.

Implementation Details. The SMC and ECD losses are implemented in C++ within the Caffe framework, as separated layers and their output added to each other and to the softmax loss. Differently from the softmax layer that is connected to the fully connected layer $fc8$, the other two losses are fed by the $pool5$ layer. The derivatives of the overall loss function (ECD contribution in Appendix A) are back-propagated by Stochastic Gradient Descend using mini-batches of 16 images and, at each iteration, the centers of each class (identity) and the sub-class (field of view) are updated accordingly. In our experiments, we integrate the losses in **ResNet50**, a state-of-the-art residual learning-based CNN formed by 53 convolutional layers [28], stacked in 16 residual blocks (Table I). Each residual block counts three convolutional layers except the first one of each *convX* aggregate that count one more as shown in Figure 9. The ResNet50 input are RGB images preprocessed by channel mean subtraction (calculated against the entire dataset) and resized to 224x224 pixels. The output of the *avg pool5* layer with dimension 2048, according to [14], is selected as deep representation of the input data. The network pretrained on the Imagenet dataset [29] is fine-tuned on the re-id datasets for the identity classification task, setting 1260 classes for CUHK03 and 751 for Market-1501, with stop point in all simulations set at 15000 iterations.

The convergence of the optimization process is regulated by a momentum $\mu = 0.9$ and a basic learning rate $\eta = 10^{-3}$ except for the last three loss layers and the fully connected layer $f8$ where we apply a learning rate multiplier of 10 in order to speed up their learning without getting to far from the original optimal point reached by pre-training. A stepwise decay policy for the learning rate is also used with dropping factor $\gamma = 0.1$ and associated step interval of 9000 iterations to progressively slow down the learning. A weight decay factor equal to $\Lambda = 5 \cdot 10^{-4}$ limits the learned network weights size. All our deep learning experiments are performed on a single machine equipped with one NVIDIA GeForce GTX Titan X GPU and an Intel Core i7-5960X CPU @3.00GHz, 64.0 GB RAM.

B. Experimented Results

State-of-the-art Methods. Our technique outperforms most of the best state-of-the-art methods in person re-identification (Table II). On CUHK03 we reach the best performance for the rank 1 accuracy (SMC+ECD, 69.55%) of all the methods listed in Table II. Also in Market-1501, SMC+ECD reports the best rank 1 accuracy (80.31%) and the best mAP (59.68%) achieving an improvement respectively of 2.1% and 10.8% of the correspondent values of the second best performing method [14]. Compared to the latter, our approach is more efficient in terms of network structure since [14] employs an architecture with two ResNet50-based branches non-sharing their weights and also performs pose estimation for generating a pose-invariant embedding and producing a confidence store to incorporate into the final data combined representation. Furthermore, we outperform the triplet loss based methods in

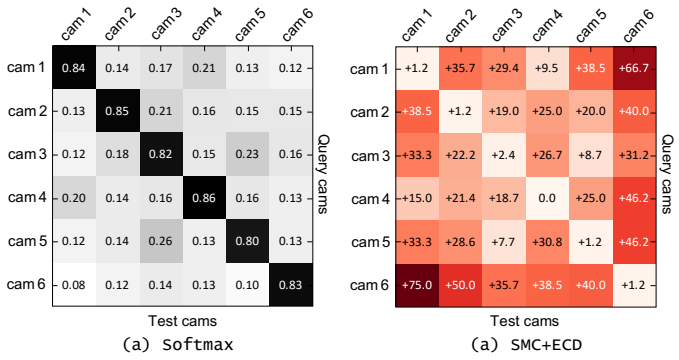


Fig. 10. Re-id performance between camera pairs on Market-1501. (a) mAP confusion matrix for the softmax baseline. (b) mAP Incremental confusion matrix (%) for SMC+ECS compared to (a).

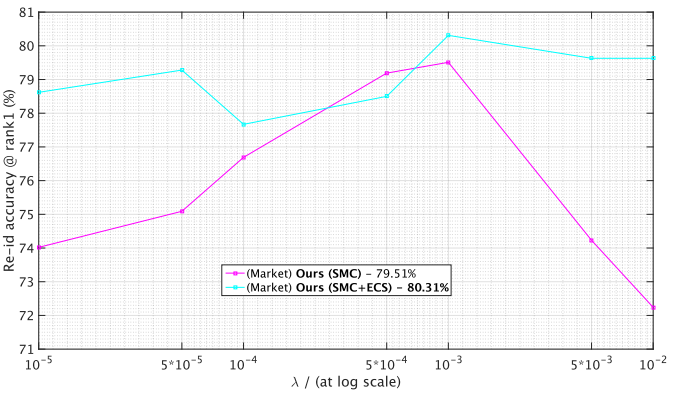


Fig. 11. Rank 1 accuracy curves to changing λ for Market-1501. λ indicates different hyper-parameters depending on the loss referred. With regards to Equation 1, when referring to the: a) SMC loss, λ denotes the λ_{SMC} parameter, with $\lambda_{ECD} = 0$; b) SMC+ECD loss, λ denotes λ_{ECD}, with $\lambda_{SMC} = 10^{-3}$. The corresponding full 2D rank 1 accuracy surface of the SMC+ECS loss, for both λ_{SMC} and λ_{ECS} changing, in Market-1501, is plotted in Figure 12.

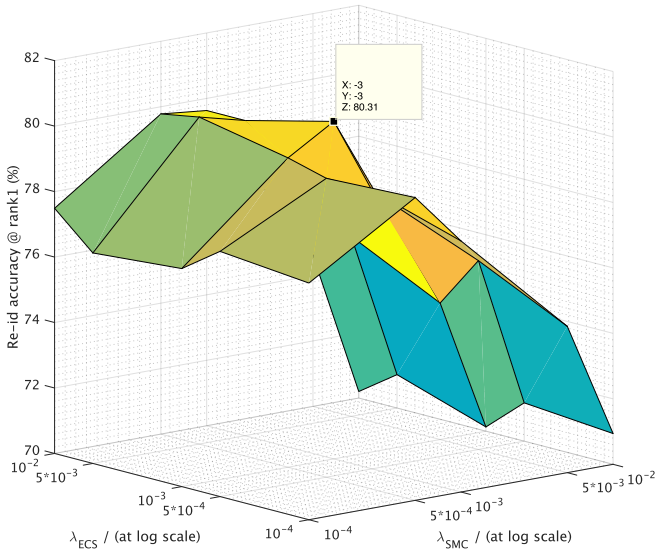


Fig. 12. Rank 1 accuracy surface of the SMC+ECD loss to changing $(\lambda_{SMC},\ \lambda_{ECD})$ hyper-parameters defined in Equation 1, for Market-1501.

[13], [51] and several techniques using Siamese networks like [8], [10], [25], [43], [44]. Our results for the multiple queries case are reported in Table IV and show further improvement.

For the sake of completeness, we point out that better performance than ours can currently be achieved on both Market-1501 and CUHK03 by a few other person re-id methods, [53]–[57]. However, these approaches should be viewed as complementary to ours more than as direct competitors, because they address aspects of the re-identification task that we intentionally exclude from the scope of our investigation. As pointed out in Section I, our approach aims to optimize the effectiveness of the training process by focusing on the loss function, under the constraint of keeping the input data, the basic training protocol and the network structure unchanged, that is by dismissing all extra sources of performance gain. Thus, we avoid to integrate in our approach strategies like data augmentation, re-ranking [56], pose/body-parts information extraction [53], [54], features fusion [53] which all take increased re-identification performance at the cost of introducing higher complexity into the training scheme/ network architecture [55]. Still, our approach can very easily be combined with them to optimize further their learning process as shown in [57] with regards to metric-learning based methods.

Ablation Analysis. Our tests performed with ResNet50 on both the single SMC loss term and the combined SMC+ECD formulation show that they outperform the single softmax loss [14], the GCL losses [21] and the center loss [20]. Our analysis has been carried out investigating the space generated by the 1D parameter λ_{SMC} for SMC and the space spanned by the 2D parameter $(\lambda_{SMC}, \lambda_{ECD})$ for SMC+ECD. Figure 11,13,14 report the rank 1 accuracy and mAP curves of SMC, SMC+ECD and also of the competing losses.

On Market-1501 SMC outperforms the softmax, GCL and center losses respectively of $+8.9\%$, $+3.8\%$, $+0.9\%$ of their rank1 accuracy and of $+22.4\%$, $+7.4\%$, $+0.9\%$ of their mAP. Likewise, for SMC+ECD, the correspondent improvements are, respectively, $+10\%$, $+4.8\%$, $+2\%$ of their rank1 accuracy and $+25.3\%$, $+10.0\%$, $+2.4\%$ of their mAP (CMC curves in Table III). On CUHK03, SMC+ECD outperforms the softmax, GCL and center loss rank 1 accuracy respectively of $+34.8\%$, $+9.3\%$, $+5.1\%$ of their value. It is noteworthy that the additional ECD loss term supervision in CUHK03 (where all identities are shot under 2 cameras only) does not provide any further improvement compared to the SMC loss term, because the ECD enforced constraint (the sum of solid lines distances in Figure 6) becomes looser and the SMC reaches by itself the maximum achievable degree of inter-class separation.

We ascertain the effectiveness of the proposed losses in the perspective of the viewpoint problem also building the mAP confusion matrix (Figure 10): it shows a significant improvement of the cross-camera re-id performance as a consequence of mitigating the effects of the viewpoint problem. The highest relative improvements are about the camera pairs $(1, 6)$ ($+66.7\%$ and $+75.0\%$), $(2, 6)$ ($+40.0\%$ and $+50.0\%$), $(5, 6)$

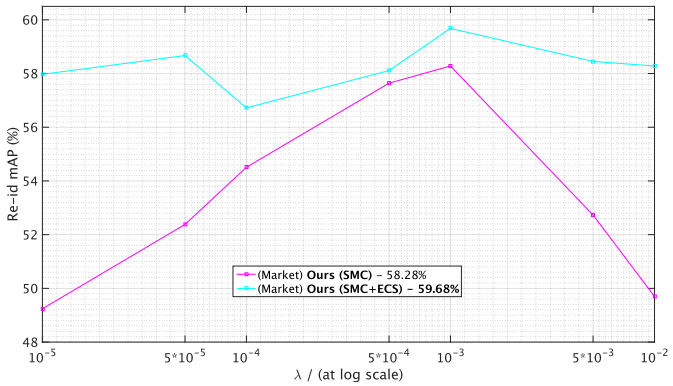


Fig. 13. mAP curves to changing λ, for Market-1501. λ indicates different hyper-parameters depending on the loss referred. With regards to Equation 1, when referring to the: a) SMC loss, λ denotes the λ_{SMC} parameter, with $\lambda_{ECD} = 0$; b) SMC+ECD loss, λ denotes λ_{ECD}, with $\lambda_{SMC} = 10^{-3}$.

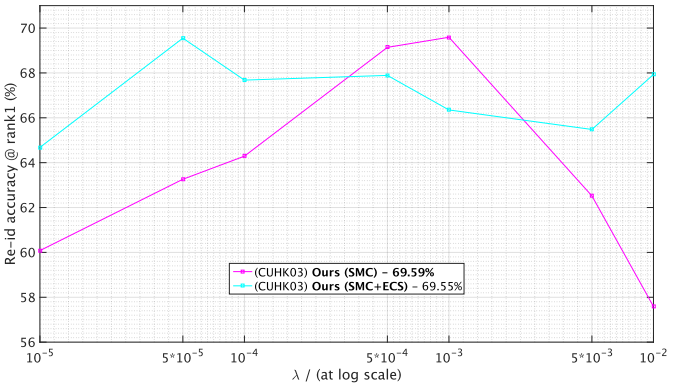


Fig. 14. Rank 1 accuracy curves to changing λ for CUHK03. λ indicates different hyper-parameters depending on the loss referred. With regards to Equation 1, when referring to the: a) SMC loss, λ denotes the λ_{SMC} parameter, with $\lambda_{ECD} = 0$; b) SMC+ECD loss, λ denotes λ_{ECD}, with $\lambda_{SMC} = 10^{-3}$.

$(+46.2\%$ and $+40.0\%)$ and $(4, 6)$ $(+46.2\%$ and $+38.5\%)$.

C. Further Results

Negatives Analysis. With regards to the analysis of the negatives of interest, Table V reports the figure of merit of all the considered losses. On Market, $F_{rank1}^{(SMC+ECD)} = 26.3\%$ and $F_{mAP}^{(SMC+ECD)} = 50.7\%$, respectively $+8.2\%$ and $+6.5\%$ higher than for the correspondent figures for the center loss. This proves the effectiveness of our loss with regards to the viewpoint problem. On CUHK03, F_{rank1} cannot be calculated because of the way its evaluation protocol [7] is defined.

Training Losses vs ML The application of the joint-Bayesian learning method to the softmax baseline shows that its performance (rank1 77.06% and mAP 53.76% in Market; rank 1 65.03% in CUHK03) is lower than that achieved by SMC+ECD without learning any metric, using the simple cosine similarity instead. This confirms that learning a similarity function in the feature space when the network weights are already fixed is sub-optimal to doing that jointly with learning the network itself, under the supervision of a more discriminative objective. Furthermore, by applying the joint-Bayesian technique to SMC and SMC+ECD there is still space for further improvement, in a measure depending on the depth and the cardinality of the dataset used ($+0.1\%$ for Market-1501 and $+3.2\%$ for CUHK03 as from Table VI).

TABLE V
IMPROVEMENT (%) OF THE "VIEWPOINT PROBLEM" IN TERMS OF RANK 1 AND MAP RELATED "FIGURES OF MERIT" IN MARKET-1501 (NOT APPLICABLE TO CUHK03).

	GCL	Center	SMC	SMC+ECD
F_{rank1}	15.5	23.4	24.3	**26.3**
F_{mAP}	33.4	35.7	47.6	**50.7**

TABLE VI
PERFORMANCE IMPROVEMENT ACHIEVED BY THE JOINT-BAYESIAN (JB) ML METHOD. APPLYING JB TO THE SOFTMAX BASELINE PERFORMS WORSE THAN PERFORMING THE TRAINING UNDER A BETTER LOSS FUNCTION SUPERVISION AND USING THE SIMPLE COSINE SIMILARITY.

		Softmax	SMC	SMC+ECD
Market-1501	**rank 1**	77.06 (+5.5)	79.93 (+0.5)	80.38 (+0.1)
	mAP	53.76 (+12.9)	58.40 (+0.2)	59.73 (+0.1)
CUHK03	**rank 1**	65.03 (+26.0)	72.04 (+3.5)	71.76 (+3.2)

V. CONCLUSION

In the context of a network of disjoint cameras, we have proposed a new loss function for supervising a CNN that were less prone to the effects of the viewpoint problem. The SMC+ECD loss represents a re-interpretation in person re-id of the center loss introduced in face verification. Furthermore, the proposed loss improves and extends the center loss by exploiting the field of view information across the datasets, critical for dealing with appearance variability. Its combined effect with the softmax loss is to simultaneously enhance the intra-class compactness (SMC) and the inter-class dispersion (ECD). Our approach outperforms most of the state-of-the-art techniques on CUHK03 and Market-1501.

ACKNOWLEDGEMENT

This work was supported in part by the UDRC consortium (University Defence Research Collaboration in Signal Processing) and Roke, Part of the Chemring Group.

APPENDIX A
LOSS DERIVATIVES

The following equations hold true:

$$\begin{aligned}\frac{\delta L_{ECD}}{\delta \boldsymbol{x}_i^{(g_i)}} = &\sum_{j=1}^{s_i}(\boldsymbol{x}_i^{(g_i)} - \boldsymbol{c}_{y_i}^{(j)}) \cdot \sum_{\substack{t=1\\t\neq i}}^{m}\sum_{k=1}^{s_t}\frac{1}{\left\|\boldsymbol{x}_i^{(g_i)} - \boldsymbol{c}_{y_t}^{(k)}\right\|_2^2} \\ &- \sum_{j=1}^{s_i}\|\boldsymbol{x}_i^{(g_i)} - \boldsymbol{c}_{y_i}^{(j)}\|_2^2 \cdot \sum_{\substack{t=1\\t\neq i}}^{m}\sum_{k=1}^{s_t}\frac{\boldsymbol{x}_i^{(g_i)} - \boldsymbol{c}_{y_t}^{(k)}}{\left\|\boldsymbol{x}_i^{(g_i)} - \boldsymbol{c}_{y_t}^{(k)}\right\|_2^4}\end{aligned} \tag{6}$$

$$\frac{\delta L_{ECD}}{\delta \boldsymbol{c}_l^{(k)}} = \begin{cases} \sum_{i=1}^{m} \left[\sum_{j=1}^{s_i} \|\boldsymbol{x}_i^{(g_i)} - \boldsymbol{c}_{y_i}^{(j)}\|_2^2 \cdot \frac{\boldsymbol{x}_i^{(g_i)} - \boldsymbol{c}_l^{(k)}}{\left\|\boldsymbol{x}_i^{(g_i)} - \boldsymbol{c}_l^{(k)}\right\|_2^4} \right], & l \neq y_i \\ \sum_{i=1}^{m} \left[(\boldsymbol{c}_l^{(k)} - \boldsymbol{x}_i^{(g_i)}) \cdot \sum_{\substack{t=1 \\ t \neq i}}^{m} \sum_{j=1}^{s_t} \frac{1}{\left\|\boldsymbol{x}_i^{(g_i)} - \boldsymbol{c}_{y_t}^{(j)}\right\|_2^2} \right], & l = y_i \end{cases} \tag{7}$$

APPENDIX B
MARKET-1501 MISLABELED IDENTITIES

By performing the negatives analysis on the top-ranked images, we have found out that the identities pairs in Table VII have been mislabeled in the original Market-1501: each pair represents a single identity. When the re-id task is addressed as a ranking problem, these 12 pairs of mislabeled images are expected to occur more than 12 times since each mislabeled pair of identities is evaluated for all their instances. This affects the performance negatively. We perform our evaluations using the original dataset, however, regardless of the mislabeling, in order to produce results fairly comparable to literature.

TABLE VII
MISLABELED IDENTITY PAIRS IN MARKET-1501.

id 1	5	13	80	157	182	198	502	746	1013	1073	1399	1446
id 2	15	1225	61	0	60	1375	1062	0	1199	91	0	675

REFERENCES

[1] W.-S. Zheng, S. Gong, and T. Xiang, "Towards open-world person re-identification by one-shot group-based verification," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 3, pp. 591–606, 2016.

[2] L. Zheng, H. Zhang, S. Sun, M. Chandraker, and Q. Tian, "Person re-identification in the wild," *arXiv preprint arXiv:1604.02531*, 2016.

[3] W. Li and X. Wang, "Locally aligned feature transforms across views," in *CVPR*, 2013.

[4] J. Chen, Y. Wang, J. Qin, L. Liu, and L. Shao, "Fast person re-identification via cross-camera semantic binary transformation," in *CVPR*, 2017.

[5] R. Zhao, W. Ouyang, and X. Wang, "Unsupervised salience learning for person re-identification," in *CVPR*, 2013, pp. 3586–3593.

[6] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Deep metric learning for person re-identification," in *ICPR*, 2014.

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

[8] R. R. Varior, B. Shuai, J. Lu, D. Xu, and G. Wang, "A siamese long short-term memory architecture for human re-identification," in *ECCV*, 2016.

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

[10] R. R. Varior, M. Haloi, and G. Wang, "Gated siamese convolutional neural network architecture for human re-identification," in *ECCV*, 2016.

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

[12] W. Chen, X. Chen, J. Zhang, and K. Huang, "A multi-task deep network for person re-identification." in *AAAI*, 2017, pp. 3988–3994.

[13] F. Wang, W. Zuo, L. Lin, D. Zhang, and L. Zhang, "Joint learning of single-image and cross-image representations for person re-identification," in *CVPR*, 2016.

[14] L. Zheng, Y. Huang, H. Lu, and Y. Yang, "Pose invariant embedding for deep person re-identification," *arXiv preprint arXiv:1701.07732*, 2017.

[15] Y.-J. Cho and K.-J. Yoon, "Improving person re-identification via pose-aware multi-shot matching," in *CVPR*, 2016.

[16] H. Shi, X. Zhu, S. Liao, Z. Lei, Y. Yang, and S. Z. Li, "Constrained deep metric learning for person re-identification," *arXiv preprint arXiv:1511.07545*, 2015.

[17] H. Shi, Y. Yang, X. Zhu, S. Liao, Z. Lei, W. Zheng, and S. Z. Li, "Embedding deep metric for person re-identification: A study against large variations," in *ECCV*, 2016.

[18] T. Xiao, H. Li, W. Ouyang, and X. Wang, "Learning deep feature representations with domain guided dropout for person re-identification," in *CVPR*, 2016.

[19] M. Geng, Y. Wang, T. Xiang, and Y. Tian, "Deep transfer learning for person re-identification," *arXiv preprint arXiv:1611.05244*, 2016.

[20] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *ECCV*, 2016.

[21] A. Borgia, Y. Hua, and N. Robertson, "A tale of two losses: Discriminative deep feature learning for person re-identification," in *IMVIP*, 2017.

[22] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[23] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *ICCV*, 2015.

[24] L. Wu, C. Shen, and A. van den Hengel, "Deep linear discriminant analysis on fisher networks: A hybrid architecture for person re-identification," *Pattern Recognition*, vol. 65, pp. 238–250, 2017.

[25] E. Ahmed, M. Jones, and T. K. Marks, "An improved deep learning architecture for person re-identification," in *CVPR*, 2015.

[26] S. Wu, Y.-C. Chen, X. Li, A.-C. Wu, J.-J. You, and W.-S. Zheng, "An enhanced deep feature representation for person re-identification," in *WACV*, 2016.

[27] A. Subramaniam, M. Chatterjee, and A. Mittal, "Deep neural networks with inexact matching for person re-identification," in *Advances in Neural Information Processing Systems*, 2016, pp. 2667–2675.

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.

[29] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[30] W. Li, X. Zhu, and S. Gong, "Person re-identification by deep joint learning of multi-loss classification," *arXiv preprint arXiv:1705.04724*, 2017.

[31] W. Liu, Y. Wen, Z. Yu, and M. Yang, "Large-margin softmax loss for convolutional neural networks," in *ICML*, 2016.

[32] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *CVPR*, 2006.

[33] S. Ding, L. Lin, G. Wang, and H. Chao, "Deep feature learning with relative distance comparison for person re-identification," *Pattern Recognition*, vol. 48, no. 10, pp. 2993–3003, 2015.

[34] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *CVPR*, 2015.

[35] E. Hoffer and N. Ailon, "Deep metric learning using triplet network," in *International Workshop on Similarity-Based Pattern Recognition*, 2015.

[36] S. Zhou, J. Wang, J. Wang, Y. Gong, and N. Zheng, "Point to set similarity based deep feature learning for person re-identification," in *CVPR*, 2017.

[37] L. Ma, X. Yang, and D. Tao, "Person re-identification over camera networks using multi-task distance metric learning," *IEEE Transactions on Image Processing*, vol. 23, no. 8, pp. 3656–3670, 2014.

[38] G. Wang, L. Lin, S. Ding, Y. Li, and Q. Wang, "Dari: Distance metric and representation integration for person verification." in *AAAI*, 2016, pp. 3611–3617.

[39] J. Hu, J. Lu, and Y.-P. Tan, "Deep transfer metric learning," in *CVPR*, 2015.

[40] D. Chen, Z. Yuan, B. Chen, and N. Zheng, "Similarity learning with spatial constraints for person re-identification," in *CVPR*, 2016.

[41] Y. Yang, S. Liao, Z. Lei, and S. Z. Li, "Large scale similarity learning using similar pairs for person verification." in *AAAI*, 2016.

[42] D. Chen, X. Cao, L. Wang, F. Wen, and J. Sun, "Bayesian face revisited: A joint formulation," in *ECCV*, 2012.

[43] E. Ustinova, Y. Ganin, and V. Lempitsky, "Multiregion bilinear convolutional neural networks for person re-identification," in *AVSS*, 2017.

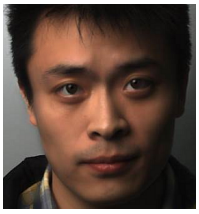
[44] L. Wu, C. Shen, and A. v. d. Hengel, "Personnet: Person re-identification with deep convolutional neural networks," *arXiv preprint arXiv:1601.07255*, 2016.

[45] J. Lin, L. Ren, J. Lu, J. Feng, and J. Zhou, "Consistent-aware deep learning for person re-identification in a camera network," in *CVPR*, 2017.

[46] L. Zhang, T. Xiang, and S. Gong, "Learning a discriminative null space for person re-identification," in *CVPR*, 2016.

[47] L. Zheng, Y. Yang, and A. G. Hauptmann, "Person re-identification: Past, present and future," *arXiv preprint arXiv:1610.02984*, 2016.
[48] S.-Z. Chen, C.-C. Guo, and J.-H. Lai, "Deep ranking for person re-identification via joint representation learning," *IEEE Transactions on Image Processing*, vol. 25, no. 5, pp. 2353–2367, 2016.
[49] B. Fernando, E. Fromont, D. Muselet, and M. Sebban, "Discriminative feature fusion for image classification," in *CVPR*, 2012.
[50] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, "Person re-identification by symmetry-driven accumulation of local features," in *CVPR*, 2010.
[51] C. Su, S. Zhang, J. Xing, W. Gao, and Q. Tian, "Deep attributes driven multi-camera person re-identification," in *ECCV*, 2016.
[52] Y. Huang, H. Sheng, Y. Zheng, and Z. Xiong, "Deepdiff: Learning deep difference features on human body parts for person re-identification," *Neurocomputing*, vol. 241, pp. 191–203, 2017.
[53] C. Su, J. Li, S. Zhang, J. Xing, W. Gao, and Q. Tian, "Pose-driven deep convolutional model for person re-identification," in *ICCV*, 2017.
[54] L. Zhao, X. Li, Y. Zhuang, and J. Wang, "Deeply-learned part-aligned representations for person re-identification," in *ICCV*, 2017.
[55] Y. Sun, L. Zheng, W. Deng, and S. Wang, "Svdnet for pedestrian retrieval," in *ICCV*, 2017.
[56] R. Yu, Z. Zhou, S. Bai, and X. Bai, "Divide and fuse: A re-ranking approach for person re-identification," *arXiv:1708.04169*, 2017.
[57] S. Bai, X. Bay, and Q. Tian, "Scalable person re-identification on supervised smoothed manifold," in *CVPR*, 2017.

Alessandro Borgia is currently a PhD student in computer vision at Heriot-Watt University and University of Edinburgh, UK. His main research interests cover machine learning and data analytics and his expertise also includes mobile network technologies and information security. He received his BSc and MSc degree in Telecommunications Engineering (2009) from La Sapienza University of Rome, Italy. He was awarded a research prize by HWU (2015) and a research contract by DSTL following an academic competition on signal processing.



Yang Hua is presently a lecturer at the Queen's University of Belfast, UK. He received his Ph.D. degree from Universit Grenoble Alpes / Inria Grenoble Rhne-Alpes, France, funded by Microsoft Research Inria Joint Center. He won PASCAL Visual Object Classes (VOC) Challenge Classification Competition in 2010, 2011 and 2012, respectively and the Thermal Imagery Visual Object Tracking (VOT-TIR) Competition in 2015. His research interests include machine learning methods for image and video understanding. He holds three US patents and one China patent.

Elyor Kodirov is currently a senior researcher at AnyVision, Belfast, UK. He received his Ph.D. degree in the School of Electronic Engineering and Computer Science, Queen Mary University of London, 2017 and the Master's degree in computer science from Chonnam National University, Korea, in 2014. His research interests include computer vision and machine learning.

Neil M. Robertson is Professor and Director of Research for Image and Vision Systems in the Centre for Data Sciences and Scalable Computing, at the Queens University of Belfast, UK. He researches underpinning machine learning methods for visual analytics. His principal research focus is face and activity recognition in video. He started his career in the UK Scientific Civil Service with DERA (2000-2002) and QinetiQ (2002-2007). Neil was the 1851 Royal Commission Fellow at Oxford University (2003-2006) in the Robotics Research Group. His autonomous systems, defence and security research is extensive including UK major research programmes and doctoral training centres.