Operator-in-the-Loop Deep Sequential Multi-camera Feature Fusion for Person Re-identification

K. L. Navaneet, Ravi Kiran Sarvadevabhatla, *Member, IEEE*, Shashank Shekhar, R. Venkatesh Babu, *Senior Member, IEEE*, and Anirban Chakraborty*, *Member, IEEE*

Abstract—Given a target image as query, person re-identification systems retrieve a ranked list of candidate matches on a per-camera basis. In deployed systems, a human operator scans these lists and labels sighted targets by touch or mouse-based selection. However, classical re-id approaches generate per-camera lists independently. Therefore, target identifications by operator in a subset of cameras cannot be utilized to improve ranking of the target in remaining set of network cameras. To address this shortcoming, we propose a novel sequential multi-camera re-id approach. The proposed approach can accommodate human operator inputs and provides early gains via a monotonic improvement in target ranking. At the heart of our approach is a fusion function which operates on deep feature representations of query and candidate matches. We formulate an optimization procedure custom-designed to incrementally improve query representation. Since existing evaluation methods cannot be directly adopted to our setting, we also propose two novel evaluation protocols. The results on two large-scale re-id datasets (Market-1501, DukeMTMC-reID) demonstrate that our multi-camera method significantly outperforms baselines and other popular feature fusion schemes. Additionally, we conduct a comparative subject-based study of human operator performance. The superior operator performance enabled by our approach makes a compelling case for its integration into deployable video-surveillance systems.

Index Terms—Person Re-identification, Surveillance, Operator-in-the-loop, Cross-camera, Feature Fusion

1 Introduction

In recent times, the development of intelligent video surveillance platforms to monitor large crowded settings such as shopping malls, railway stations, airports etc. has become a priority to ensure public safety and security. A crucial component of such a platform is the *person re-identification* (re-id) system. Given a query image, a re-id system searches through all the camera Field-of-Views (FoVs) and returns a per-camera ranked list of candidate matches. However, due to large variation in illumination, viewpoint, target resolution and other challenges arising from occluded targets, re-id methods are often unable to retrieve the correct match within a short enough ranked list. This imposes a significant burden on human operators of the surveillance system who now need to laboriously scan large lists per camera. The problem is further compounded when a large number of cameras are present. Such factors have kept person re-id an open problem in computer vision.

In a deployment scenario, it is fairly typical to observe a person in more than one camera FoV. Since each observation may provide complementary information, the human operator must seek the target in *every* per-camera ranked list generated by a re-id system. If the target is identified in a particular list, the operator may choose to 'label' the same via a simple haptic operation (e.g. touch or mouse-based selection). However, in a classical re-id scheme, the per-camera lists are generated *independently* [1, 2, 3] without

* *Corresponding author*
The authors are with the Department of Computational and Data Sciences, Indian Institute of Science, Bangalore, India, 560012.
E-mail: navaneetl@iisc.ac.in, ravika@gmail.com, shashankshek@iisc.ac.in, venky@iisc.ac.in, anirban@iisc.ac.in
This paper has supplementary downloadable materials including additional details and results available at http://ieeexplore.ieee.org.

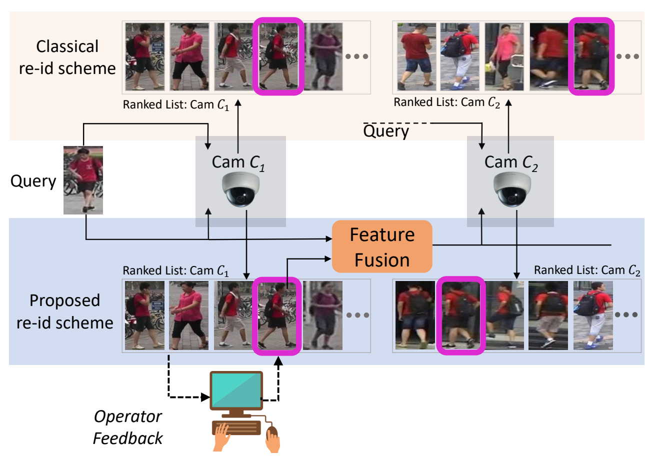

Fig. 1: (Top) Classical re-id scheme where query image's feature representation is used to search each camera in the network independently. The retrieved lists are returned to the human operator. (Bottom) A real-world deployment scenario that motivates our proposed sequential re-id scheme where operator feedback regarding target sighting is utilized towards better re-id performance in an online fashion. In the figure, camera C_1 is queried first and ranked list of matches is obtained. The correct match (pink box) in the retrieved list is identified by operator and is subsequently fused with query image at feature level (orange block). This fused representation is used to query camera C_2. Notice that ranking of query target in C_2's list is expected to improve in the sequential fusion-based approach unlike the classical version which cannot exploit operator inputs to improve subsequent queries.

taking actions of the human operator into account. In other words, target labeling by the operator in a subset of cameras cannot be leveraged to improve the ranking of the query target in the remaining set of cameras (see 'Classical re-id scheme' in Figure 1).

It is certainly desirable to exploit the complementary information on target appearance from multiple camera FoVs and consequent operator labeling. To this end, we propose a novel sequential and iterative approach which improves ranking of the target as additional cameras are queried across the network. Towards the success of our approach, we develop a sequential multi-camera fusion scheme. The fusion scheme operates on feature representations of candidate matches (see 'Proposed re-id scheme' in Figure 1). Our approach has three major advantages. Firstly, it can accommodate an arbitrary number of cameras. Secondly, the fusion scheme is flexible enough to operate on cameras in any arbitrary order. Thirdly and crucially, our approach is designed to produce a *monotonic* improvement in re-id performance as additional target labels from different cameras are fused.

In addition, the proposed approach naturally aligns with the manner in which a human operator typically interacts with a re-id system. Therefore, it can be seamlessly integrated into deployable video-surveillance systems. The proposed approach is also designed as plug-and-play, i.e., it can be used atop any state-of-the-art camera pairwise feature estimation/metric learning method for re-id. Therefore, improvements in the camera-pairwise re-id approaches can be utilized and further extended within our framework. Concretely, we make the following contributions:

- We propose a novel framework for utilizing feedback from human operators in a re-id pipeline deployed in a real-world scenario. In this proposed framework, observations from query target in a subset of cameras can be aggregated to obtain improved retrieval results for the remaining cameras in the network (Sec. 3).
- We propose a novel sequential feature fusion scheme and a training strategy that learns to achieve monotonic improvement in re-id performance as additional observations from the target are fused. (Sec. 3.4).
- To demonstrate the effectiveness of our approach, we define novel test protocols (Sec. 4.3) and perform extensive experiments (Sec. 4.4) on two large-scale multi-camera benchmark datasets (Market-1501 [4], DukeMTMC-reID [5]).
- We perform comparative analysis of human operator performance obtained from interaction logs of a deployed re-id user interface to demonstrate the superiority and real-world feasibility of our approach.

We define two novel protocols to evaluate our proposed fusion framework. While both these protocols are directly motivated from the deployment scenario described in Fig. 1, they are also carefully modified to enable quantitative evaluation of fusion as well as comparison with traditional re-id approaches.

2 Related Work

The problem of person re-identification has been well studied over the last decade [6]. An important class of person re-id methods involve development of feature descriptions that are discriminative between different targets and exhibit robustness to variations in viewpoint, color, illumination etc. across different camera FoVs [7, 8, 9, 10, 11, 12, 13]. Popular discriminative signature-based methods include ICT [14], SDALF [15], saliency based methods [16, 17] , hierarchical Gaussian descriptors [18] and many more. Besides these, a large volume of works have focused on camera-pairwise metric learning techniques [19, 20, 21, 22]. Some widely used such techniques are LADF [23], RankSVM [24], KISSME [1], LFDA [25], CFML [21] and XQDA [2].

Recently, deep neural network based person re-id approaches have shown significant performance improvements by jointly learning the feature representation and the distance metric [26, 27, 28, 29, 30, 31, 32, 33]. Unlike the classical hand-crafted techniques where the feature extraction and the metric learning methods were independently designed and cascaded, deep learning approaches jointly optimize for these two interconnected components, outperforming the *non-deep* methods in the process. Many such methods solve re-id as a verification/binary classification problem. A popular approach involves Siamese networks with contrastive loss [3, 34]. In [26], LSTM modules were introduced into a Siamese network to model spatial dependencies between image parts. [35] proposed a domain-guided dropout strategy to make the learned re-id model robust to inter-dataset variations. Even beyond Siamese, [36] provides an improved triplet loss for obtaining a more discriminative feature representation. In datasets with large number of unique identities [4], robust feature representations can be learned in an identification mode, i.e., training to map each image to an ID and using the learned feature embedding to associate unseen IDs during testing phase [6, 37, 38]. Specifically, in [6], the authors implemented a modified ResNet-50 [39] model on Market-1501 [4] dataset under both identification and verification setup. We adopt the verification based protocol and baseline model in our experiments.

Recurrent Neural Networks have been used for feature aggregation in various video-based applications[40, 41, 42]. Feature fusion for person re-id has also been considered, but in a multi-query set-up where multiple images of a target from the same camera are fused using simple pooling operations on feature representations [4]. Multi-camera fusion has been employed for object detection [43], tracking [44] and activity classification [45]. While there are works in other fields with operator/human-in-the-loop frameworks, they essentially differ from our work in the manner in which the human feedback is made use of. For e.g, [46] tries to learn similarity between face images from probe and gallery sets with human assistance. The work uses similarity labels as feedback from humans to iteratively embed the query into the learned feature set. Similarly, multi-camera feature fusion has been considered in the literature, but in a way unlike the proposed approach. Images from multiple cameras are used either in training or during inference to obtain a single fused representation which is then used for decision making ([47, 48]). In contrast, the proposed fusion framework involves sequential fusion of inputs from multiple cameras. The current fused representation is used to query and retrieve images from gallery set and the retrieval features are then combined with the existing fused feature to obtain the subsequent query. To the best of our knowledge, ours is the first work to utilize operator feedback in such a sequential framework to perform fusion at the feature level.

3 Proposed Approach

In this section, we lay out details of our method. We begin with a formal problem statement of our fusion approach (Section 3.1). Having done so, we identify three key properties that need to be satisfied during the fusion process (Section 3.2). We subsequently

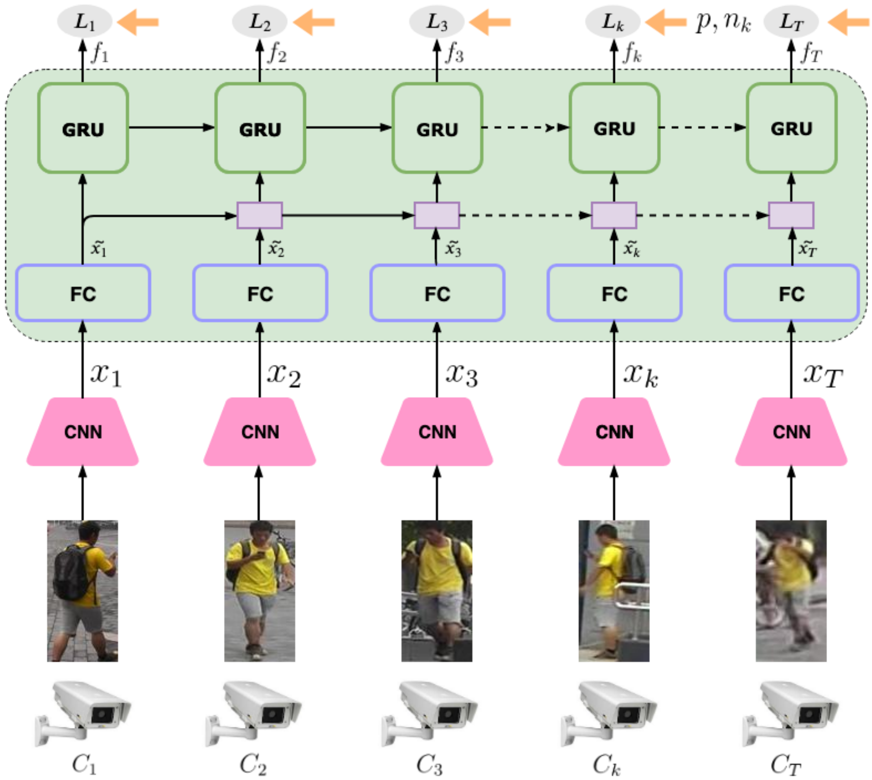

Fig. 2: An illustration of our fusion architecture (Sec. 3). The baseline CNN features $(x_1, x_2, \ldots)$ from camera images are fed to our fusion function. Purple boxes indicate mean-pooling of corresponding inputs. The fusion network is optimized via a novel loss formulation L_k, applied at each time step k, to improve the accumulated feature representation (f_k). p and n_t are the representations for positive and negative instances. Note that for a given training sequence, the person id (anchor) and positive instance are held constant across the cameras, while negative instances vary.

present the design of the fusion function (Sec. 3.3) and our novel modifications to the default optimization procedure (Sec. 3.4), all designed to satisfy the properties mentioned previously. A GRU based implementation of the proposed fusion function is described in Sec. 3.5.

3.1 Problem Statement

Obtaining discriminative person-specific representations is a key component of any modern re-id approach. To obtain such representations, we follow the standard convention of fine-tuning pre-trained Convolutional Neural Networks (CNNs) on person re-id datasets for classification/verification task. For a given person image, we use the corresponding final, fully-connected layer's output of the fine-tuned CNN as the feature representation and employ x or its subscripted variants to refer to the same.

Our problem can now be stated as follows: Suppose the total number of cameras is T and the human operator has performed selection of the query target in $k \leqslant T$ cameras. Given the sequence of corresponding features $\{x_1, x_2, \ldots, x_k\}$, the aim is to learn a fusion function $\mathcal{F}$ that integrates operator feedback and produces an optimal fused representation f_k, i.e. $f_k = \mathcal{F}(x_1, x_2, \ldots, x_k)$[∗].

3.2 Desired Properties of the Fusion Function

The number of camera FoVs in which a query is visible can vary from target to target. Therefore, the fusion function $\mathcal{F}$ must be capable of handling a variable number of input feature representations. In addition, images of the same target in different camera FoVs often provide complementary visual information. Hence, a proper fusion of these image features should produce a more robust and holistic feature representation that leads to a better re-id accuracy/mAP. To achieve these aims, the proposed fusion approach must ideally satisfy the following properties:

1) $\mathcal{F}$ must be able to process camera (feature) sequences of variable lengths, i.e. k can vary from target to target.
2) As the number (k) of feature representations being aggregated increases, the fused representation f_k should improve, i.e. enable sustenance or increase in re-id accuracy.
3) $\mathcal{F}$ should be invariant to relative ordering in the input feature sequence, i.e. the order in which cameras are considered should not matter.

3.3 Design of the Fusion Function

A feature fusion module can be designed in a number of possible ways. Among the popular early fusion/feature fusion techniques, mean and max pooling (element-wise for the feature vectors) can be suitable candidates for our fusion function as both of these satisfy the desired properties-1 and 3 by design. However, these methods do not necessarily guarantee the property-2, i.e., the fused representations resulting in sustenance or improvement in re-id accuracy when longer sequences of features are input to the fusion function. Towards this, the function should be designed such that it contains learnable parameters and the desired properties (e.g., property-2) can be implicitly enforced via minimization of

∗. Please note the distinction between fixed feature representations (x) obtained from CNN and the 'learnt' fused representations (f) produced by our fusion function.

a suitable cost over these parameters. Most recurrent models (e.g. recurrent neural nets) would be classified under this category of functions. In the current and the following subsection, we describe the design of the suitable cost functions for optimal estimation of the fusion function parameters.

During the training phase, we require $\mathcal{F}$ to transform the sequence of image features $\{x_1, x_2, \ldots, x_k\}$ from the k different cameras to a corresponding sequence of fused representations $\{f_1, f_2, \ldots, f_k\}$ (Sec. 3.1). To achieve this, the image features are first transformed to an embedding of pre-defined dimension to obtain $\{\tilde{x}_1, \tilde{x}_2, \ldots\}$. To increase the robustness of the fusion process, $\{\tilde{x}_1, \tilde{x}_2, \ldots\}$ up to and including current camera index t are mean-pooled (purple boxes in Fig. 2) and fed as input to a recurrent function block.

Suppose we choose an image from a training sequence and define it as the *anchor*. We define *positive* instances as those training images having the same id as that of the anchor and *negative* instances as those images whose id differs from anchor's id. Ideally, we require that a positive instance's feature representation be closer to anchor's representation than the negative's representation.

This objective can be achieved via minimization of a hinge-style *triplet* loss [36, 49, 50, 51] defined on the anchor, positive and negative instance representations.

$$L^{(tri)} = \sum_{\{f,p,n\}} \max(0, \|f - p\|_2 - \|f - n\|_2 + m) \quad (1)$$

where m is the margin.

In our setting, we set up the triplet loss $L_t^{(tri)}$ for each camera index t wherein the fused representation f_t serves as the anchor. The choice of positive instances is limited, being confined to same sequence or at the most a handful of other sequences. We omit the camera corresponding to the positive instance during the fusion process since images from the same camera have high similarity. We also choose to keep the positive instance fixed for *all* indices of a given training sequence. The negative instance for each index is chosen using hard mining within a given training mini-batch [51]. To enable comparison with the fused feature, the positive and negative instances are processed by the fusion module $\mathcal{F}$ for a single time-step to obtain the corresponding features p and n_t. The triplet loss at index t is defined as:

$$L_t^{(tri)} = \sum_{\{f_t,p,n_t\}} \max(0, \|f_t - p\|_2 - \|f_t - n_t\|_2 + m) \quad (2)$$

We use soft-margin formulation as an approximation to the hinge loss [51] as follows:

$$L_t^{(tri)} = \sum_{\{f_t,p,n_t\}} \ln(1 + e^{\|f_t - p\|_2 - \|f_t - n_t\|_2}) \quad (3)$$

3.4 Monotonic Representation Improvement

To specifically address the requirement of progressive improvement in the quality of the fused representation f_t (property-2 in Section 3.2), we introduce an additional per-index loss term called *monotonicity loss* (m-loss). m-loss is formulated as a sum of zero-margin hinge losses as follows:

$$L_t^{(mon)} = \textstyle\sum_{\{f_t,p,n_t\}} \max(0,\ d(f_t,p) - d_p^{*(t)}) + \max(0,\ d_n^{*(t)} - d(f_t,n_t)) \quad (4)$$

where $d(.)$ is the euclidean distance metric. $d_p^{*(t)}$ and $d_n^{*(t)}$ are defined as follows:

$$d_p^{*(t)} = \min_{\tau \in \{1,2,\ldots,t-1\}} d(f_\tau, p) \quad (5a)$$

$$d_n^{*(t)} = \max_{\tau \in \{1,2,\ldots,t-1\}} d(f_\tau, n_\tau) \quad (5b)$$

Eq. 4 and eq. 5(a) ensure that the fused representation at step t is closer to the positive instance than all the fused representations till index $t-1$. Also, the negative instance is chosen using hard-mining within a mini-batch and $d_n^{*(t)}$ is chosen as the maximum of distances from the fused representations to the corresponding negative samples. Therefore eq. 4 and eq. 5(b) enforce f_t to be farther from all negative samples in the mini-batch compared to any fused representation till step $t-1$.

The total loss at each time step t is formulated as a weighted combination of the triplet loss and the monotonicity loss, i.e.

$$L_t = L_t^{(tri)} + \lambda \lambda_t^R L_t^{(mon)} \quad (6)$$

While λ is fixed for all indices t, λ_t^R is obtained using a linear weighting scheme to give more importance to monotonicity loss for longer sequences. For a sequence of length T, $\lambda_t^R = t/T$. Overall, the proposed loss formulation is designed to ensure a decoupled optimization of the two desired properties – low triplet loss when a new feature representation is aggregated and monotonic improvement in fused feature representation.

3.5 Implementation of the Fusion Module

To meet the requirements for the fusion function as described above, we judiciously design $\mathcal{F}$ around as a recurrent neural network. Specifically, out of many choices (RNNs, LSTMs, GRUs etc.) for the recurrent architectures, we choose to use a Gated Recurrent Unit (GRU) (Fig. 2) [52] - a popular Recurrent Neural Network architecture (Sec. 3.3). In GRU, the following set of transformations are applied at each index t of the sequence:

$$r_t = \sigma(W_{rx} x_t + W_{rh} h_{t-1} + b_r) \quad (7a)$$

$$z_t = \sigma(W_{zx} x_t + W_{zh} h_{t-1} + b_z) \quad (7b)$$

$$s_t = \tanh(W_{hx} x_t + W_{hh}(h_{t-1} \odot r_t) + b_h) \quad (7c)$$

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot s_t \quad (7d)$$

Here, $\odot$ represents element-wise multiplication and σ represents the sigmoid function. h_t is formulated to serve as an effective feature representation for the input feature sequence $\{x_1, x_2, \ldots, x_t\}$ seen until that point, i.e., $f_t = h_t$. The intermediate transformations r_t, z_t, s_t are formulated such that the GRU effectively fuses only helpful aspects of the input and ignores the rest. Our design choice of GRU is significantly motivated by this property. Note that the subscripted W's and b's are shared across all the sequence indices and form the trainable parameters of the GRU.

3.6 Training and Testing

The sequence-loss for GRU is computed as an average across per-index total loss (eq. 6). During the fusion network training, we nominally fix an input camera sequence ordering and the inputs to the GRU are obtained on the basis of this ordering. We emphasize that the choice of ordering is arbitrary. In fact, we shall show later that the camera ordering has negligible effect on re-id performance

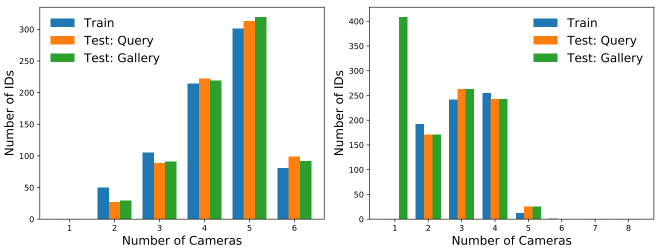

Fig. 3: Histogram of maximum number of cameras each target is observed in Market-1501 [4] (left) and DukeMTMC-reID [53] (right) datasets. In DukeMTMC-reID, there are very few samples with query sequence lengths greater than 4.

(Sec. 4.4). This result also implies that the fusion function satisfies the third property from the desirable properties of an ideal fusion function (Sec 3.2).

In the testing phase, query images from multiple cameras are considered for fusion. We use the hidden state h_k of the GRU at the last camera index (Eq. 7(d)) as the fused feature f_k. Since the ids of images in the gallery set are unknown, it is not possible to obtain a fused representation for them. To enable comparison between query and gallery features, we construct a sequence by repeating the gallery image and use it as the input to the GRU. Additional details on this procedure are presented in Sec. 4.3.

Other Fusion Functions: We explore *mean-pooling* and *max-pooling* of features as two alternative fusion functions. As discussed earlier in this section, both these functions (with no trainable parameters) satisfy the desired properties 1 and 3 by design. These pooling operations are performed in ways similar to multi query setting for person re-identification [4] to obtain the fused representations. We present a detailed comparative evaluation of the fusion functions in Sec. 4. We also show how the early/feature fusion based sequential re-id compares in performance with two late-fusion approaches (Sec. 4.4.4).

4 Experiments

4.1 Datasets

Since the focus of the work is on fusion of features from multiple cameras, we evaluate performance on datasets with a minimum of three cameras in the network. We report our results on two such datasets, Market-1501 and DukeMTMC-ReID, which contain 6 and 8 cameras respectively.

Market-1501 [4]: This dataset has $12{,}936$ images from 751 IDs in the training set and another 750 test IDs with $3{,}368$ and $19{,}732$ images in the query and gallery sets respectively. Each ID is present in a minimum of two and a maximum of six cameras (see left plot in Fig. 3). The gallery set has multiple instances of an ID from a camera while the query set has only one. All the images are of dimensions 128×64.

DukeMTMC-ReID [53]: This dataset is organized similar to Market-1501. It has 702 IDs each in the train and test sets. There are $16{,}522$, $2{,}228$ and $17{,}661$ images in train, query and gallery sets respectively. All the images are obtained using manually annotated bounding boxes. In the training set, each ID is present in a minimum of 2 and a maximum of 6 cameras, even though the network has 8 cameras (Fig. 3). The gallery set has 408 distractor IDs, not present in more than one camera FoV.

4.2 Implementation Details

4.2.1 Feature Extraction

For our experiments, we use ResNet-50 [39] and AlexNet [54] as the base (per camera image) CNN feature extractor models. Note that these choices are nominal and any off-the-shelf model can be used as the baseline feature extractor.

For the ResNet-50 baseline, we use the network pre-trained on ImageNet [55] for fine-tuning on reID datasets. An additional fully-connected (FC) layer is used at the end of Pool-5 layer of ResNet-50 to reduce the feature dimension to 512. For the AlexNet baseline, we remove Local Response Normalization and employ batch-normalization at every layer before the non-linearity. Similar to the ResNet-50 set-up, the output embedding dimension is set to 512. During the baseline network training, dropout with rate 0.5 is employed for the fully-connected layers. We use Adam optimizer with an initial learning rate of 0.0001. β_1 and β_2 parameters in the optimizer are set to 0.9 and 0.999 in all experiments. As done in [51], the learning rate is decreased as the training progresses according to the following schedule:

$$\epsilon(t) = \begin{cases} \epsilon_0 & \text{if } t \leqslant t_0 \\ \epsilon_0 \times 0.001^{\left(\dfrac{t-t_0}{t_1-t_0}\right)} & \text{if } t_0 \leqslant t \leqslant t_1 \end{cases} \tag{8}$$

Here, ϵ_0, t_0 and t_1 are set to 0.0001, 15000 and 25000 respectively.

The input dimensions for ResNet-50 and AlexNet are fixed to 256×128 and 227×227 respectively and the input images are accordingly resized. To maintain the aspect ratio of input in ResNet-50, the pooling layer is modified to enable an input of dimension 256×128. Following [54], we augment our training set with 5 random crops and their mirrored images. The size of crop is set to 89% of the original image size.

4.2.2 Fusion Function

The GRU is initialized with random weights and hidden state length is set to 512 in all our experiments. As in CNN training, we use the Adam optimizer to perform gradient descent. For the experiments with monotonicity loss (Sec. 3.4), the weighting factor λ is calculated using the scheduling scheme similar to that in Eq. 8 (ϵ, ϵ_0 replaced with λ, λ_0) with λ_0 equal to 0.01.

4.3 Evaluation Protocols

In the protocol generally followed for evaluation in multi-camera setting [6], single query and single gallery sets are used irrespective of number of cameras in the network. The images from all the cameras are binned together in the gallery and for a given query, predictions from the same camera are treated as inadmissible, i.e. not considered for evaluation. In our work, we tackle the novel task of cross-camera fusion which requires a minimum of two camera inputs into the fusion function and at least one gallery camera to compare the fused representations against. This setting is different from traditional protocols and hence the existing evaluation procedures cannot be directly adopted.

Therefore, we modify the traditional protocols under the constraints present in re-id datasets and propose two new evaluation protocols – Variable Set Protocol (VSP) and Fixed Set Protocol (FSP). These two protocols are explained in detail in the following sections. The suitability aspect towards evaluation of our proposed

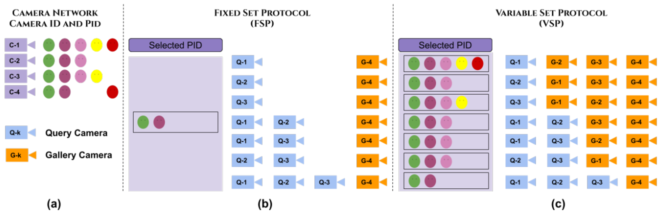

Fig. 4: Comparison of camera subset and ID selection in FSP and VSP evaluation protocols. An illustration of the protocols is provided for a sample camera network with four cameras and five unique identities, denoted by different colored circles. Different query-gallery camera subsets and corresponding IDs for querying are shown for each protocol. Note that for VSP, only a subset of all possible camera combinations are shown for brevity.

framework and design justifications for each of these protocols are also discussed in detail.

4.3.1 Variable Set Protocol (VSP):

Note that we require a comparison of the proposed approach with traditional re-id methods used as baselines in this work (along with alternative fusion approaches). Therefore, we develop a protocol characteristically very similar to the traditional re-id evaluation setups, while suitably modified to align with our sequential re-id philosophy. For this, we partition the dataset into two sets of observations, which are not only mutually exclusive in terms of their image contents, but are also disjoint in terms of camera field-of-views (FoVs) from which the observations are sourced.

Let C be the set of cameras present in the network (Fig. 4 a). A subset of C is considered as the gallery camera set $\{G_C\}$. The complementary set of $\{G_C\}$ is considered to be the query camera set $\{Q_C\}$. For evaluation, a set of query person IDs from $\{Q_C\}$ are selected such that they are present in a minimum of one camera in $\{G_C\}$ (Fig. 4 c). This procedure is repeated for all possible gallery camera sets. Finally results are averaged over query subsets of same cardinality. The total number of such query-gallery combinations is given by $N = \sum_{i=1}^{n-1} \binom{n}{i} = 2^n - 2$ where n is the number of cameras in the network. Note that the size and contents of both the query and gallery sets change based on the selection of number of cameras for fusion (each cluster of bars in Fig. 5). Hence, it is not possible to compare the performance of fusion function for different lengths of sequences across these different sets of partitions, following this protocol. We have designed the FSP protocol specifically to study this.

4.3.2 Fixed Set Protocol (FSP):

We design this protocol to specifically evaluate the fusion approaches. In an example deployment setup, shown in Fig. 1, the retrieved image from a camera is fused with the query and a *different* camera with non-overlapping field-of-view is queried to obtain the subsequent retrieval. This procedure is repeated to obtain fusions of growing number of images until all the cameras in the network are exhausted. Thus the query set before and after a certain fusion step changes in size and contents. To evaluate the proposed fusion scheme, one needs to freeze the gallery set to allow for a fair comparison of fusion performance across different query subsets. This practical constraint of having the same evaluation setting for comparing different lengths of fusion compels us to design an alternative evaluation protocol, termed 'Fixed Set Protocol' (FSP). Instead of fusing query images with the retrieved images from gallery, images of the same target from the query camera subsets are used for fusion and then evaluation of the performance of this fusion is performed on the fixed gallery set ($\{G_C\}$). Note that the inputs to be fused are selected using ground truth labels, and can be thought of as the 'virtual' human operator's feedback. To evaluate the fusion function's performance, we consider all possible subsets of cameras within the query set ($\{Q_C\}$), starting with one camera (no fusion) and progressively increasing until N_Q, the number of cameras within the query set. For further clarifying these motivations as well as towards better understanding of this protocol, refer to Fig. 4 (a,b). Let there be four cameras in a network, numbered 1, 2, 3, 4. Cameras 1, 2 and 3 are query cameras, fused in this particular order and the camera 4 constitutes the gallery. In the first step, query from camera 1 is combined with a 'retrieved' (we use ground truth labels to simulate retrieval by human operator) image from camera 2 to query the gallery set. Subsequently input from camera 3 is combined with the previous fused representation to again query the same gallery set. This procedure is repeated for all possible query camera combinations. The total number of such possible query camera combinations in any camera network is $N = |\mathcal{P}(Q_C)| - 1$ where $\mathcal{P}(S)$ and $|S|$ are the power set and the cardinality of $\{S\}$ respectively. Note that we choose only those IDs which are present in all the cameras in both $\{Q_C\}$ and $\{G_C\}$ so as to enable fusion in any query camera subset (Fig. 4). Thus the set of query IDs is *fixed* for a given gallery set regardless of the query subset used, rendering the metrics for different query subset combinations comparable.

In the test phase for both protocols, feature fusion is performed only on the query subset of the dataset. To enable comparison of query and gallery features during testing, we mimic the multi-camera scenario by constructing a sequence of repeated gallery image features. Our decision is motivated by the fact that our fusion function is optimized for sequences and also by better performance observed in practice. We empirically set the number of gallery image repetitions to be same as the query sequence length. In the following sub-sections, we report results for the GRU based fusion function trained only with triplet loss (termed GRU) in both VSP and FSP protocols. To show the efficacy of m-loss, we would need to compare performance across different query sequence lengths and thus report results for GRU trained with both triplet and m-loss (termed GRU+m-loss) only on the FSP protocol.

Overall, the proposed VSP and FSP protocols enable us to evaluate a realistic deployment scenario and quantitatively compare such a scenario with traditional baseline schemes. More specifically, FSP has been designed to compare the utility of fusion and the proposed 'm-loss' (Section 3.4) across variable-length observation sequences. In contrast, VSP is aimed broadly towards comparison of the proposed GRU based fusion framework with traditional re-id models used as baselines in this work.

4.4 Results

4.4.1 Baseline CNN performance

Table 1 shows the rank-1 and mAP metrics for the ResNet-50 and AlexNet CNN baselines on Market-1501 and DukeMTMC-ReID datasets. The ResNet based network significantly outperforms the AlexNet based network. The above pre-trained baseline networks are used as the feature extractors for the fusion module in all our

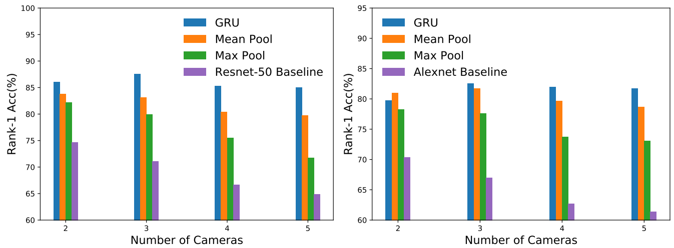

Fig. 5: Rank-1 accuracies for Variable Set Protocol on Market-1501 dataset with ResNet-50 (left) and AlexNet (right) CNN baselines. GRU based fusion outperforms mean and max-pooling based fusion methods. All fusion techniques are significantly better than the CNN baselines.

Architecture	Market-1501		DukeMTMC-reID	
	Rank-1 ↑	mAP ↑	Rank-1 ↑	mAP ↑
ResNet-50	73.63	48.74	60.86	39.79
AlexNet	67.1	44.34	56.87	34.21

TABLE 1: Classification-based Baseline CNN performance.

Architecture	mAP for FSP			mAP for VSP		
	Proposed	Mean-pool	Max-pool	Proposed	Mean-pool	Max-pool
ResNet-50+Fusion	**75.87**	67.88	64.52	**71.82**	64.65	61.29
AlexNet+Fusion	**67.79**	67.13	63.65	**63.71**	63.10	59.44

TABLE 2: Comparison of averaged mAP on Market-1501 with unit length galleries. The proposed fusion methodology consistently outperforms the other feature fusion techniques. Note that the baseline CNN performance for VSP is 54.31% and 51.05% respectively for ResNet-50 and Alexnet based architectures.

experiments. Since ResNet based network achieves better retrieval performance, we primarily show results using the ResNet baseline.

4.4.2 Results with VSP

The results for VSP on Market-1501 are shown in Fig. 5. Since the baseline feature extractor methods take in inputs from only one camera at a time, we independently query from each of the cameras in the query set. The scores are computed for each of these individual queries and their average is considered for comparison with feature fusion based methods. In this protocol, we report results for GRU based fusion function trained with just the triplet loss.

For better representation, we average the results based on the number of cameras present in the query set. From the results (Fig. 5), we observe that our approach (fusion of queries) performs significantly better than baseline – for ResNet-50, on average, fusion outperforms baseline by $\mathbf{13.5}\%$ and mean-pool based fusion by $\mathbf{3.6}\%$ in Rank-1 accuracy. The mAP performances (Table 2) are more noteworthy with $\mathbf{17.5}\%$ and $\mathbf{7.2}\%$ improvement over baseline CNN and mean-pool based fusion respectively. In the case of AlexNet as baseline CNN, mean-pool based fusion performs slightly better than our approach for sequences of length two. However, as the number of cameras increase, our approach outperforms all other approaches, thereby satisfying a design objective of our fusion function. Additionally, the figures show that the improvement obtained using feature fusion increases as more query cameras are considered, as expected.

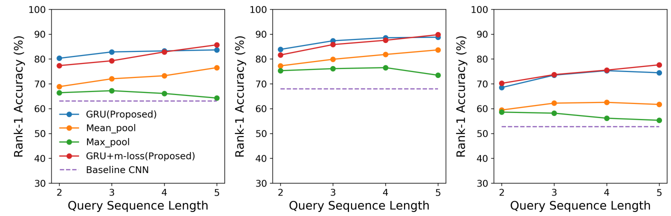

Fig. 6: Effect of query sequence lengths on accuracy. Rank-1 accuracies on three different gallery sets are shown for fusion with ResNet-50 baseline on Market-1501 dataset. Feature fusion using GRU performs significantly better than other fusion techniques in rank-1 accuracies, while m-loss helps in maintaining monotonicity with increasing sequence lengths

Architecture	Rank-1 @ Query Sequence Length			
	2	3	4	5
ResNet-50+Fusion	79.37	83.58	84.29	85.01
AlexNet+Fusion	71.21	76.88	80.14	81.77

TABLE 3: Comparison of averaged FSP rank-1 accuracies of ResNet-50 and AlexNet based fusion with varying query set lengths. Evaluation is performed on the GRU+m-loss based fusion. Results are averaged over all six unit length gallery camera sets.

4.4.3 Results with FSP

To show the efficacy of fusion, we compare fusion performance for varying query sequence lengths with fixed gallery sets. The query sequence length refers to the cardinality of the query camera combination. In this protocol, we compare both the GRU (triplet only) as well as the same trained using the additional m-loss to show the utility of the latter as the length of sequence of observations to be fused increases. Table 3 presents the comparison of the proposed fusion based re-id using ResNet-50 and AlexNet baselines on Market-1501 dataset using FSP for different query sequence lengths. The gallery camera set length is fixed to one. Hence, at most five images can be used for feature fusion. The average rank-1 accuracies over six such galleries is shown in the table. ResNet-50 based fusion network performs significantly better due to better baseline features. Hence, in the remaining experiments on FSP, we present results mainly on ResNet-50 architecture. The effect of number of query images on fusion accuracy can also be viewed in Fig. 6. The monotonic trend of accuracies with increase in number of query cameras holds in the case of GRU alone, but is further enhanced when trained with m-loss, leading to improved accuracy at the later time-steps. On an average, our fusion approach achieves $\mathbf{5.8}\%$ improvement in Rank-1 accuracy over mean-pooling. Table 2 provides a comparison of mAP with ResNet-50 and AlexNet architectures on Market-1501. For ResNet-50, our fusion approach outperforms mean-pool based fusion in mAP by about $\mathbf{8}\%$. The significant improvement in mAP indicates that the fused representation is able to effectively combine images, leading to better low-rank retrievals. The results also crucially highlight the advantage of our GRU-based fusion over simple pooling approaches. Fig. 7 presents averaged FSP results on gallery sets with two cameras on Market-1501 dataset. Since there are two cameras in the gallery set, the maximum possible query sequence length is four. As in the case of length one gallery sets, we observe a monotonic improvement in the retrieval

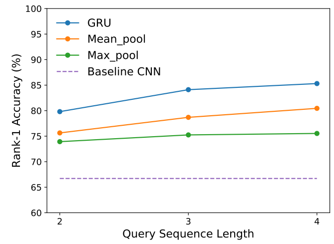

Fig. 7: Averaged FSP results for size 2 gallery sets using ResNet-50 based network on Market-1501 [4] dataset. The fusion based approaches result in monotonic improvement with increase in query sequence length. The proposed GRU based fusion significantly outperforms the other fusion techniques.

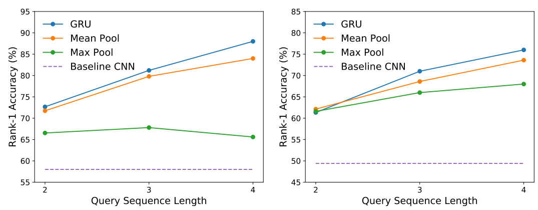

Fig. 8: Rank-1 accuracy for Fixed Set Protocol on DukeMTMC-reID dataset with ResNet-50 (left) and AlexNet (right) CNN baselines.

performance of all fusion methodologies as more images are fused. In summary, the proposed GRU based fusion techniques with and without m-loss significantly outperform the baseline fusion approaches and the effect is pronounced with increasing query sequence lengths, especially, when m-loss is additionally imposed while training the GRU.

Fig. 8 presents rank-1 accuracy results on the DukeMTMC-reID dataset following the FSP protocol. Due to dearth of query sequences with length greater than four, we consider query sets with a maximum of four cameras, while gallery size is fixed to two. The results are averaged over all such possible gallery sets. Our approach consistently outperforms other fusion techniques on both ResNet-50 and AlexNet baselines, while increasing the accuracy with fusion. We provide additional results on a third dataset (MSMT17 [56]) in Section 4 of supplementary.

4.4.4 Comparison with Late Fusion Baselines

In sections 4.4.2 and 4.4.3, we compared the proposed GRU based fusion with other early fusion schemes, namely mean and max pooling. We further substantiate our choice of fusion function through comparison with late fusion based approaches too. Specifically, we use the features from the baseline CNN and perform score fusion and maximum probability based fusion. In both these schemes, fusion is done on the distances between query-gallery image features rather than at the feature level. That is, Euclidean distances between query and gallery image features are calculated independently for individual input features to be fused, which are subsequently combined using a weighted average to obtain the final distance values after fusion. In score fusion, equal weight is given to each of the input features to be aggregated. In maximum probability fusion, a discrete probability distribution over the gallery set is obtained by normalizing the distance of the query from the gallery images. The distance corresponding to the query input having maximum probability for a given gallery image is considered to be the fused distance. Quantitative results under both FSP and VSP protocols on the Market-1501 dataset are given in Table 4. In the case of FSP, we observe that the performance of the late fusion techniques are similar to that of other baseline mean pool based fusion scheme and better than the max-pool scheme. In VSP, both the late fusion schemes are significantly better than mean/max pool for all query sequence lengths. However, the proposed GRU based fusion scheme consistently outperforms all the other baseline fusion schemes (early/late fusion) by a large margin.

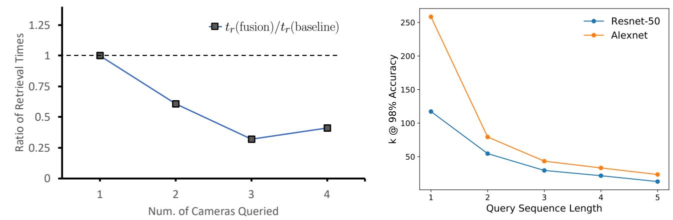

Fig. 9: Ratio of human operator retrieval times with and without fusion (left) and automated retrieval list length (right) plots on Market-1501 dataset.

Approach	Protocol	Rank-1 @ Query Sequence Length			
		2	3	4	5
GRU+m-loss	FSP	**79.37**	**83.58**	**84.29**	**85.01**
Score Fusion		73.60	77.30	79.11	80.44
Max Prob. Fusion		69.92	73.25	75.75	78.06
GRU	VSP	**86.03**	**87.54**	**85.31**	**85.01**
Score Fusion		84.80	83.78	81.17	80.58
Max Prob. Fusion		80.96	79.50	77.24	78.19

TABLE 4: Rank-1 accuracy comparison with late fusion approaches on Market-1501 dataset. Proposed GRU based fusion scheme consistently outperforms all the other baseline fusion schemes by a large margin.

4.4.5 Portability of the Fusion Scheme

In the proposed fusion framework, the fusion function training is independent of the choice of feature extraction pipeline. The feature extraction network parameters are not updated during the training of fusion network. Though we choose a ResNet-50 based model trained on camera-pairwise re-identification tasks, the framework can easily accommodate any other general feature extraction pipeline. This plug-and-play nature of the proposed pipeline would enable us to seamlessly integrate any feature extractor that is used in traditional re-id setup, and the overall retrieval performance would surely benefit from any progress in the classical/traditional re-id. To further substantiate this claim, we show retrieval results of the proposed framework atop a state-of-the-art conventional person re-id approach. Specifically, we use the pre-trained model of HA-CNN [57] to obtain the image feature representations as input to the GRU based fusion function. HA-CNN learns soft attention at the pixel level and hard attention at the region level and improves the feature representation through the use of a 'harmonious attention' module. The training of our proposed fusion module with HA-CNN as feature extractor is done in a manner identical to that explained in the previous sections and the retrieval results are shown on Market-1501 dataset in Table 5. We observe that the results are consistent with that obtained

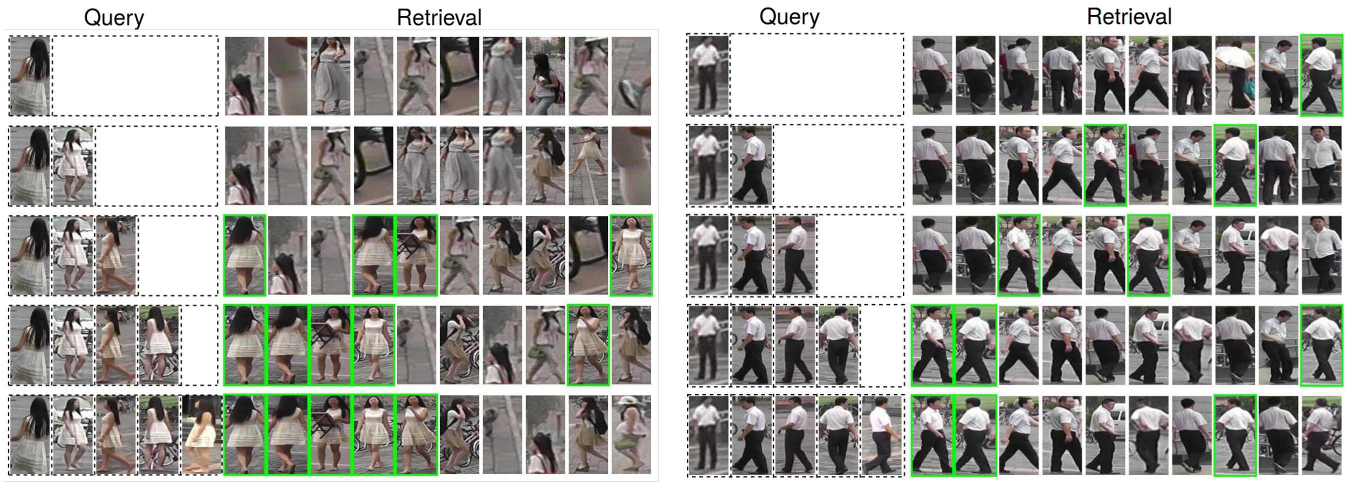

Fig. 10: Retrieved samples for two example targets from Market-1501 dataset. Correct retrievals are indicated with green box. More correct matches are obtained at a lower rank as additional query images are combined (best viewed in color).

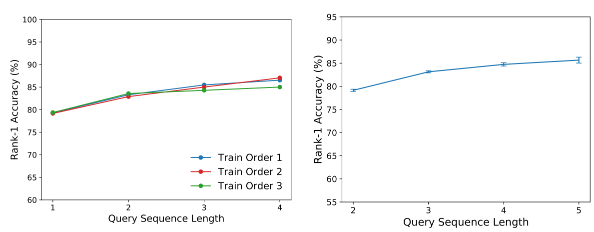

Fig. 11: Averaged FSP results for different input ordering sequences during training (left) and testing (right).

Approach	Protocol	Rank-1 @ Query Sequence Length			
		2	3	4	5
GRU+m-loss	FSP	**83.65**	**85.69**	**86.13**	**87.50**
HA-CNN Baseline		78.72			
GRU	VSP	**91.24**	**92.18**	**90.72**	**88.59**
HA-CNN Baseline		91.03	89.06	85.63	83.58

TABLE 5: Portability of the proposed fusion framework: We observe that fusion improves the retrieval accuracy when used atop HA-CNN [57], a state-of-the-art re-id feature extraction pipeline.

using the ResNet-50 baseline, i.e., the proposed re-id framework with the GRU fusion scheme achieving impressive improvements in retrieval performance over the baseline HA-CNN based re-id across both the FSP and VSP protocols. Also, it can be noted that the superiority of the fusion framework is apparent even in scenarios where the baseline achieves high retrieval accuracy.

4.4.6 Advantages of Fusion in Deployed Systems

To study the performance advantages of employing fusion-based algorithms in practical surveillance systems, we designed a prototype GUI system (Fig. 6 in supplementary materials) for human-operator-in-the-loop re-id and conducted a comparative user study to determine the relative time spent in retrieval with and without the fusion of queries.

We showed 15 different identities on an average to a pool of subjects recruited for the study. In the GUI, the query image is displayed on the left and the corresponding top-k retrievals are displayed in the right panel in the order of increasing ranks (Fig. 6 in supplementary). We display 25 retrievals ($k = 25$) per page on the GUI. The subject searches through the retrievals and selects the matching image. If the subject is unable to find the right match, the next k (25) retrievals are displayed. This process continues until the subject successfully locates a match. The retrieved image is then fused with the query to obtain retrievals in the subsequently queried camera. A similar experiment is performed without fusion, i.e., by querying each camera independently with one single image or retrieved target image from the preceding camera (without fusion). In Fig. 9 (left), we plot the ratio of the average time taken for retrieval with and without fusion (t_r(fusion)/t_r(baseline)) as a function of query sequence lengths (i.e., the number of cameras queried). We observe that retrieval times are significantly smaller and decrease with increasing query sequence length with our fusion-based approach in contrast to the conventional approach involving independent querying, thereby reinforcing the practical utility of the proposed framework.

The average rank of first correct retrieval (termed 'minimum retrieval list length') as obtained by our algorithm is shown in Fig. 9 (right). The retrieval list length decreases monotonically with query sequence length, emphasizing the advantage of proposed approach.

In Fig. 10, we present two sample sequences of queries and corresponding top-10 retrievals. As the fusion function processes more images, the number of correct retrievals within top-10 ranks increases. Fusion is especially beneficial in challenging scenarios where multiple candidates with near-identical appearances exist in the gallery with minute differences between them (Fig. 10 (right)). Note that, while more correct retrievals are obtained within top-10 ranks as images are fused, there is an improvement in the position (rank) of the existing retrievals too. This indicates that our approach is able to integrate new information while retaining the relevant aspects of the existing representation.

4.4.7 Effect of Camera Ordering

As discussed in Sec. 3, we desire the fusion function to be agnostic to input ordering in both the training and testing phases. To verify this, we train the fusion network with multiple sequence orders corresponding to different camera arrangements. We observe that the average FSP results on six unit length galleries are similar across training orders (Fig. 11 (left)). Conversely, for a fixed training order, we examined multiple orderings of query cameras during testing. We sample 50 randomly ordered sequences of length 5 and according to FSP (4.3.2), consider all possible combinations of sub-sequences for each sequence. The mean rank-1 accuracy and the standard deviations are plotted in Fig. 11 (right). As can be seen (Fig. 11 (right)), the fusion performance is practically independent of camera ordering in this case as well.

5 Conclusion

In this paper, we have proposed a novel sequential multi-camera feature fusion approach for person re-id. Unlike classical re-id methods, our approach can accommodate operator inputs in an online fashion, enabling early gains via a monotonic improvement in target retrieval accuracy. These capabilities are made possible by our choice of GRU as a fusion function and our training strategy involving a custom formulation of the monotonicity loss. We also introduce novel evaluation protocols and conduct extensive evaluations on Market-1501 and DukeMTMC-reID datasets. The results indicate that our multi-camera fusion method significantly outperforms the corresponding baselines as well as other popular feature fusion schemes. Additionally, our comparative analysis of operator-in-the-loop performance showcases the potential for seamless integration into deployable video-surveillance systems.

Zheng et al. [58] proposed a temporal metric for evaluation of re-id systems in a temporally changing dynamic gallery set scenario. It would be interesting to examine the connections between the temporal metric of Zheng et al. and the VSP protocol proposed in our current work since both deal with variable gallery sets. The current version of our work is not designed to explicitly omit noisy/spurious features from a camera, especially during the testing phase. One possibility would be to incorporate attention mechanisms in future to accomplish the same and further improve fusion during both training and testing phases.

6 Acknowledgement

This work is partially supported by Pratiksha Trust, Bangalore and Robert Bosch Centre for Cyber Physical Systems, IISc.

References

[1] M. Koestinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," in *CVPR*. IEEE, 2012, pp. 2288–2295. 1, 2

[2] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *CVPR*, 2015, pp. 2197–2206. 1, 2

[3] E. Ahmed, M. Jones, and T. K. Marks, "An improved deep learning architecture for person re-identification," in *CVPR*, 2015, pp. 3908–3916. 1, 2

[4] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *ICCV*, 2015, pp. 1116–1124. 2, 5, 8

[5] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in *ECCV Wksp.*, 2016. 2

[6] L. Zheng, Y. Yang, and A. G. Hauptmann, "Person re-identification: Past, present and future," *arXiv preprint arXiv:1610.02984*, 2016. 2, 5

[7] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," *Computer Vision–ECCV 2008*, pp. 262–275, 2008. 2

[8] I. Kviatkovsky, A. Adam, and E. Rivlin, "Color invariants for person reidentification," *PAMI*, vol. 35, no. 7, pp. 1622–1634, 2013. 2

[9] N. Martinel and C. Micheloni, "Re-identify people in wide area camera network," in *CVPR(W),*. IEEE, 2012, pp. 31–36. 2

[10] C. Liu, S. Gong, C. C. Loy, and X. Lin, "Person re-identification: What features are important?" in *European Conference on Computer Vision*. Springer, 2012, pp. 391–401. 2

[11] Y. Hu, S. Liao, Z. Lei, D. Yi, and S. Li, "Exploring structural information and fusing multiple features for person re-identification," in *CVPR Workshops*, 2013, pp. 794–799. 2

[12] R. Zhao, W. Ouyang, and X. Wang, "Learning mid-level filters for person re-identification," in *CVPR*, 2014, pp. 144–151. 2

[13] D. S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino, "Custom pictorial structures for re-identification." in *BMVC*, vol. 2, no. 5, 2011, p. 6. 2

[14] T. Avraham, I. Gurvich, M. Lindenbaum, and S. Markovitch, "Learning implicit transfer for person re-identification," in *ECCV 2012 Workshops*, 2012, pp. 381–390. 2

[15] L. Bazzani, M. Cristani, and V. Murino, "Symmetry-driven accumulation of local features for human characterization and re-identification," *CVIU*, vol. 117, no. 2, pp. 130–144, 2013. 2

[16] R. Zhao, W. Ouyang, and X. Wang, "Person re-identification by salience matching," in *ICCV*, 2013, pp. 2528–2535. 2

[17] ——, "Unsupervised salience learning for person re-identification," in *CVPR*, 2013, pp. 3586–3593. 2

[18] T. Matsukawa, T. Okabe, E. Suzuki, and Y. Sato, "Hierarchical gaussian descriptor for person re-identification," in *CVPR*, 2016, pp. 1363–1372. 2

[19] K. Q. Weinberger and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," *JMLR*, vol. 10, no. Feb, pp. 207–244, 2009. 2

[20] M. Hirzer, P. M. Roth, M. Köstinger, and H. Bischof, "Relaxed pairwise learned metric for person re-identification," in *ECCV*. Springer, 2012, pp. 780–793. 2

[21] B. Alipanahi, M. Biggs, A. Ghodsi *et al.*, "Distance metric learning vs. fisher discriminant analysis," in *Natl. Conf. on AI*, vol. 2, 2008, pp. 598–603. 2

[22] M. Dikmen, E. Akbas, T. S. Huang, and N. Ahuja, "Pedestrian recognition with a learned metric," in *ACCV*, 2010. 2

[23] Z. Li, S. Chang, F. Liang, T. S. Huang, L. Cao, and J. R. Smith, "Learning locally-adaptive decision functions for person verification," in *CVPR*, 2013, pp. 3610–3617. 2

[24] B. J. Prosser, W.-S. Zheng, S. Gong, T. Xiang, and Q. Mary, "Person re-identification by support vector ranking." in *British Machine Vision Conference*, vol. 2, no. 5, 2010, p. 6. 2

[25] S. Pedagadi, J. Orwell, S. Velastin, and B. Boghossian, "Local fisher discriminant analysis for pedestrian re-identification," in *CVPR*, 2013, pp. 3318–3325. 2

[26] 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*. Springer, 2016, pp. 135–153. 2

[27] H. Liu, J. Feng, M. Qi, J. Jiang, and S. Yan, "End-to-end comparative attention networks for person re-identification," *IEEE Transactions on Image Processing*, 2017. 2

[28] C. Su, S. Zhang, J. Xing, W. Gao, and Q. Tian, "Deep attributes driven multi-camera person re-identification," in *ECCV*. Springer, 2016, pp. 475–491. 2

[29] W. Chen, X. Chen, J. Zhang, and K. Huang, "A multi-task deep network for person re-identification." in *AAAI*, 2017. 2

[30] D. Li, X. Chen, Z. Zhang, and K. Huang, "Learning deep context-aware features over body and latent parts for person re-identification," in *CVPR*, 2017, pp. 384–393. 2

[31] 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, pp. 1288–1296. 2

[32] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Deep metric learning for person re-identification," in *ICPR*. IEEE, 2014, pp. 34–39. 2

[33] W. Li, R. Zhao, T. Xiao, and X. Wang, "Deepreid: Deep filter pairing neural network for person re-identification," in *CVPR*, 2014, pp. 152–159. 2

[34] R. R. Varior, M. Haloi, and G. Wang, "Gated siamese convolutional neural network architecture for human re-identification,"

in *ECCV*. Springer, 2016, pp. 791–808. 2

[35] T. Xiao, H. Li, W. Ouyang, and X. Wang, "Learning deep feature representations with domain guided dropout for person re-identification," in *CVPR*, 2016, pp. 1249–1258. 2

[36] 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, pp. 1335–1344. 2, 4

[37] Z. Zheng, L. Zheng, and Y. Yang, "A discriminatively learned cnn embedding for person re-identification," *arXiv preprint arXiv:1611.05666*, 2016. 2

[38] Y. Sun, L. Zheng, W. Deng, and S. Wang, "Svdnet for pedestrian retrieval," in *ICCV*. IEEE, 2017, pp. 3820–3828. 2

[39] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778. 2, 5

[40] N. McLaughlin, J. Martinez del Rincon, and P. Miller, "Recurrent convolutional network for video-based person re-identification," in *CVPR*, 2016, pp. 1325–1334. 2

[41] S. Xu, Y. Cheng, K. Gu, Y. Yang, S. Chang, and P. Zhou, "Jointly attentive spatial-temporal pooling networks for video-based person re-identification," in *CVPR*, 2017. 2

[42] J. Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, "Beyond short snippets: Deep networks for video classification," in *CVPR*, 2015. 2

[43] S. Bhinge, Y. Levin-Schwartz, and T. Adalı, "Data-driven fusion of multi-camera video sequences: Application to abandoned object detection," in *ICASSP*, 2017, pp. 1697–1701. 2

[44] S. L. Dockstader and A. M. Tekalp, "Multiple camera fusion for multi-object tracking," in *Multi-Object Tracking, 2001. Proceedings. 2001 IEEE Wksp. on*. IEEE, 2001, pp. 95–102. 2

[45] M. Hekmat, Z. Mousavi, and H. Aghajan, "Multi-view feature fusion for activity classification," in *ICDSC*. ACM, 2016. 2

[46] C. Cao and H.-Z. Ai, "Facial similarity learning with humans in the loop," *Journal of Computer Science and Technology*, vol. 30, no. 3, pp. 499–510, May 2015. 2

[47] J. Harguess, C. Hu, and J. Aggarwal, "Fusing face recognition from multiple cameras," in *2009 Workshop on Applications of Computer Vision (WACV)*. IEEE, 2009, pp. 1–7. 2

[48] E. Ristani and C. Tomasi, "Features for multi-target multi-camera tracking and re-identification," in *CVPR*, 2018. 2

[49] 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. 4

[50] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *CVPR*, 2015, pp. 815–823. 4

[51] A. Hermans, L. Bayer, and B. Leibe, "In Defense of the Triplet Loss for Person Re-Identification," *arXiv preprint arXiv:1703.07737*, 2017. 4, 5

[52] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Gated feedback recurrent neural networks," in *ICML*, 2015, pp. 2067–2075. 4

[53] Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled samples generated by gan improve the person re-identification baseline in vitro," in *ICCV*, 2017, pp. 3754–3762. 5

[54] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012, pp. 1097–1105. 5

[55] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *CVPR*. IEEE, 2009, pp. 248–255. 5

[56] L. Wei, S. Zhang, W. Gao, and Q. Tian, "Person transfer gan to bridge domain gap for person re-identification," in *CVPR*, 2018, pp. 79–88. 8

[57] W. Li, X. Zhu, and S. Gong, "Harmonious attention network for person re-identification," in *CVPR*, June 2018. 8, 9

[58] M. Zheng, S. Karanam, and R. J. Radke, "Measuring the temporal behavior of real-world person re-identification," *arXiv preprint arXiv:1808.05499*, 2018. 10

K.L. Navaneet is a masters student in Department of Computational and Data Sciences at Indian Institute of Science (IISc), Bangalore, India. He received his bachelors degree from Department of Electrical and Electronics Engineering, National Institute of Technology-Karnataka, Surathkal, India. His research interests include computer vision, machine learning and image processing.

Ravi Kiran Sarvadevabhatla is an Assistant Professor at International Institute of Information Technology-Hyderabad (IIIT-H). He received his Ph.D. from Dept. of Computational and Data Sciences, Indian Institute of Science, Bangalore, India. He has broad-ranging research interests and likes to work on problems involving multi-modal multimedia data (e.g. images, videos, text, audio/speech, eye-tracking data) and multiple disciplines (e.g. Humanities, Graphics, Robotics, Human-Computer Interaction).

Shashank Shekhar is a Project Assistant at Visual Computing Lab, Department of Computational And Data Sciences, Indian Institute Of Science, Bangalore, India. He received his B.Tech. in Electronics And Communication Engineering from Indian Institute of Technology (ISM) Dhanbad, India in 2017 before working as a Software Engineer at Samsung Research Institute, Delhi for a while. His research interests include computer vision, representation learning and reinforcement learning.

R. Venkatesh Babu is an Associate Professor in Department of Computational and Data Sciences, Indian Institute of Science, Bangalore. He received his Ph.D. degree from the Dept. of Electrical Engineering, Indian Institute of Science. Thereafter, he held postdoctoral positions at NTNU, Norway and IRISA/INRIA, Rennes, France. Subsequently he worked as a research fellow at NTU, Singapore. His research interests span signal processing, compression, machine vision, image/video processing, pattern recognition and multimedia. He is a senior member of IEEE.

Anirban Chakraborty is an Assistant Professor in the Department of Computational and Data Sciences, Indian Institute of Science, Bangalore. He received his Ph.D. in Electrical Engineering from the University of California, Riverside in 2014. Subsequently, he held research fellow positions with the National University of Singapore and Nanyang Technological University. Anirban's research interests lie in the broad areas of computer vision, machine learning, optimization etc. and their applications in problems such as data association over large graphs, data fusion, video surveillance, video-based biometrics, bio-image informatics etc. He is a member of IEEE.