Attention-based Few-Shot Person Re-identification Using Meta Learning

Alireza Rahimpour Hairong Qi
Department of Electrical Engineering and Computer Science
University of Tennessee, Knoxville, TN, USA
{arahimpo,hqi}@utk.edu

Abstract

In this paper, we investigate the challenging task of person re-identification from a new perspective and propose an end-to-end attention-based architecture for few-shot re-identification through meta-learning. The motivation for this task lies in the fact that humans, can usually identify another person after just seeing that given person a few times (or even once) by attending to their memory. On the other hand, the unique nature of the person re-identification problem, i.e., only few examples exist per identity and new identities always appearing during testing, calls for a few-shot learning architecture with the capacity of handling new identities. Hence, we frame the problem within a meta-learning setting, where a neural network based 'meta-learner' is trained to optimize a 'learner' — an attention-based matching function. Another challenge of the person re-identification problem is the small inter-class difference between different identities and large intra-class difference of the same identity. In order to increase the discriminative power of the model, we propose a new attention-based feature encoding scheme that takes into account the critical intra-view and cross-view relationship of images. We refer to the proposed Attention-based Re-identification Meta-learning model as ARM. Extensive evaluations demonstrate the advantages of ARM as compared to the state-of-the-art on the challenging PRID2011, CUHK01, CUHK03 and Market1501 datasets.

1. Introduction

Recently, person re-identification has gained increasing research interest in the computer vision community due to its importance in multi-camera surveillance systems. In person re-identification, the goal is matching people across non-overlapping camera views at different times. Despite all the research efforts, person re-identification remains a challenging problem since a person's appearance can vary significantly when large variations in view angle, illumination, human pose, background clutter and occlusion are involved, as shown in Fig. 1. To address these difficulties, several approaches have been proposed in recent years. These algorithms either focus on learning more discriminative metrics for comparing person images or extracting more representative visual features. Specifically, inspired by the success of deep neural networks in many computer vision tasks, deep architectures have been widely used for person re-identification and achieved state-of-the-art results (e.g., [13, 63, 1, 78]). However, there are still challenging issues remaining in solving the person re-identification problem.

Figure 1: Two examples of camera view variation challenge. The images of the same person (the ones with green border in each row) look different in two different camera views. Also, images from different people in the gallery may look the same. Our model is able to handle these challenging situations.

The first challenge is the lack of examples per identity as well as difficulties in training these deep learning based models. Most recent supervised deep learning based approaches consider re-identification as a classification problem and have demonstrated better performance [85] than the matching-based models (e.g., Siamese [76]). However, these methods need lots of labeled training data per ID which might not be feasible for real-world deployment with data exhibiting the characteristic of having many different classes but few examples per class. In fact, in most of the existing re-identification datasets the number of images per identity for each view is on average less than 5 (e.g., each

identity has 2, 4.8 and 3.6 images on average for each view in CUHK01, CUHK02 and Market1501, respectively).

Another trend is to consider re-identification as a matching problem which would address the lack of example issue. Recent development in this direction includes the triplet loss and its variants [13, 57, 16, 66]. However, training of these triplet-based networks is challenging and requires specific algorithms (e.g., hard negative mining [53]) for selecting the triplets and the margin. Improper selection can in practice lead to bad local minima early on in training or very slow convergence. In addition, since the triplet loss based networks consider the pairwise relationship between images as triplets only during training, they suffer from an imperfect generalization capability in testing [10]. Finally, semi-supervised and unsupervised methods have been proposed [37, 83] to address the need for labeled data, but unfortunately their performance has not been quite in par with supervised methods.

The second challenge is how to model the relationship between all the gallery and probe images in the feature extraction process (in both training and testing). In fact, different individuals can share a similar appearance, and the appearance of the same person can be drastically different in different camera views. Fig. 1 shows two instances of view variations challenge. For example, even though the gallery images of the second example (b) look the same, they actually belong to different identities. Therefore, cross-view and intra-view relationships play a critical role in person re-identification. Most existing approaches fail to define a structure that compares the probe image to the whole gallery images as a set when extracting the feature representation of each image. The performance of learned features using existing criterion based on pairwise similarity is still limited, because only point-to-point (i.e., image-to-image) relationships are mostly considered in a Siamese or triplet loss structure. Although classification-based approaches use a softmax layer where the relationship between images is implicitly considered, the feature of each image is still extracted independently of all the other images in the gallery.

To **address the first challenge**, we propose an attention-based model for few-shot person re-identification. The motivation lies in the fact that humans, even children, can usually generalize after seeing the same person only a couple of times. Moreover, few-shot re-identification would help alleviate data collection and annotation in large camera networks as we would not require large amount of labeled examples to attain reasonable performance. The proposed model is able to learn quickly from a few examples during training and is able to identify new people in the gallery in the testing stage. In order to deal with new classes in the gallery and to acquire quick knowledge from few examples during training, we exploit the idea of meta-learning. Meta-learning suggests framing the learning problem at two levels. The first level is the 'learner' that learns a metric for performing the matching between each probe image and the gallery images. The second level is the 'meta-learner' that guides the 'learner' across all the matching processes. Unlike previous works, by using this framework, there is no need for any additional procedure for triplet or margin selection during training. Furthermore, since meta-learning optimizes an objective during training which directly reflects the person re-identification task during testing, the model has more generalization power to cope with unobserved identities without any changes to the network.

To **address the second challenge**, we propose an attention-based feature encoding structure, where we take into account the relationship among all the gallery images as well as that between all the gallery and probe images in the feature encoding process. The relationship between all the images inside the gallery is modeled using the proposed gallery encoder architecture. Similarly, the relationship between the probe image and the gallery set is modeled using the probe encoder architecture. The proposed feature encoding framework enhances the discriminative capability of feature representation by leveraging the cross-view and intra-view relationship between images and selectively propagating relevant contextual information throughout the network. In this way, compared to existing works, the proposed network can deal with the view-specific matching task more effectively. Moreover, the memory structure in the attention-based encoding framework helps the meta-learning algorithm to remember how to update the learner, thus facilitates the handling of new unseen identities in the gallery. We refer to the proposed Attention-based Re-identification Meta-learning model as ARM. The flowchart of ARM is shown in Fig. 2.

The main contributions of this paper are summarized as follows:

- Introducing a novel end-to-end attention-based person re-identification framework which is able to perform few-shot learning by exploiting the concept of meta-learning.
- Designing an attention-based gallery encoder to incorporate the intra-view and cross-view relationships between gallery images in order to generate more representative and discriminative features.
- Designing an attention-based probe encoder in order to model the cross-view relationship between the probe and gallery images. The proposed task-driven encoder attends to the encoded features of all the gallery images and incorporates their inter-relationship into feature representation of each probe image.

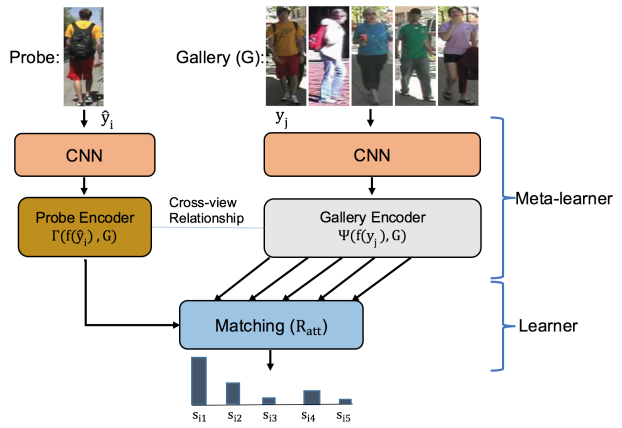

Figure 2: Flowchart of the proposed Attention-based Re-identification Meta-learning model (ARM) in training and testing. Given a probe image (top left), the model learns to attend to its best match in the gallery set (top right). The attention-based re-identification function R_{att} calculates the probability s_{ij} of each gallery image y_j (i.e., in this toy example: $j = 1, \ldots, 5$) to be a match for the probe image $\hat{y_i}$ (Best viewed in color).

2. Related Work

In general, existing approaches for person re-identification are mainly focused on two aspects: learning a distance metric [32, 33, 42, 56, 73] and developing a new feature representation [41, 82, 84, 10]. In distance metric learning methods, the goal is to learn a metric that emphasizes inter-personal distance and de-emphasizes intra-personal distance. The learnt metric is used to make the final decision as to whether a person has been correctly re-identified or not. In the second group of methods based on developing new feature representation for person re-identification, novel feature representations have been proposed to address the challenges such as variations in illumination, pose and view-point [63, 34, 82, 84].

In recent years, several person re-identification approaches based on deep neural network architecture have been proposed and achieved outstanding results [78, 13, 30, 1, 66, 71, 61, 85, 80, 86, 28]. In most of the Convolutional Neural Network (CNN)-based methods for re-identification, the goal is to jointly learn the best feature representation and a distance metric. Furthermore, a new trend has been started recently for designing a more realistic end-to-end re-identification framework based on joint detection and identification of people from video frames (i.e., person search problem) [85, 72]. Recent advances in Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) models [17, 77, 59] provide some insights as to how to integrate the contextual information in the model. It has been shown in [43] that LSTM cells can detect salient keywords relevant to a topic (i.e., context) from sentences or speech inputs. In [62], a Siamese LSTM architecture was presented that can process image regions sequentially and enhance the discriminative capability of local feature representation.

With the recent surge of interest in deep neural networks, attention-based models have been shown to achieve promising results on several challenging tasks [75, 4, 3]. Very recently a few works have considered the importance of attention in the re-identification task [45, 35, 87]. In [35] and [45] a spatial attention mechanism was employed in order to detect saliency in the images in a triplet loss structure. In [87] a temporal attention model was added to the spatial attention model in a Siamese framework [8] in order to select the informative frames over the sequence of frames in the video. Different from these existing works, in our proposed model we employ LSTM-based attention in order to model the intra-view and cross-view relationships between images in a novel feature encoding framework.

The other related works to our proposed model are the re-identification methods based on one-shot learning [6] and the methods that investigate the relationship between the gallery images [86, 20]. However, the proposed method in [6] is a different one-shot metric learning framework from ours and the main idea was to infer a similarity metric on unknown testing domain (e.g. camera pair) by having access to a single pair of images from this camera pair. In [20], the relationship between the images in the gallery was studied based on a re-ranking framework for removing the visual ambiguities common to first ranks.

Different from these related works and inspired by the exciting recent developments based on meta-learning [64, 50], attention mechanism [4, 65], and memory augmented neural networks [51, 70, 25, 2] in this paper, we study the impact of the *attention mechanism* and *meta-learning* in solving the few-shot person re-identification problem. Our work uses the meta-learning framework introduced in [64]. However in [64], the model performance is sensitive to the order of examples in the support set (i.e., memory) since a bi-directional LSTM has been used for feature embedding. That is, the bi-directional LSTM gives different embeddings for different ordered support sets. However in our framework this problem has been solved thanks to the content-based attention encoding which is not sensitive to the order of images in the gallery and also leads to lower computational complexity of the model. Furthermore, we use improved distance metrics for the attention models which showed better re-identification performance. Our system is end-to-end trainable using just a few examples of each identity and is able to generate representative and discriminative features by using the proposed feature encoding architecture.

3. Method

In this section, we first elaborate on the proposed attention-based feature encoder that incorporates intra-view and cross-view relationships in the feature extraction process. This is followed by detailed explanation of the attention-based matching model. The proposed meta-learning based few-shot re-identification and its training and testing procedure are discussed in the end.

3.1. Attention-based Feature Encoder

Learning discriminative feature representation is one of the key objectives of our re-identification model. Changes in camera views usually generate significant appearance changes due to the variations in view angle, background, illumination and occlusion. Therefore, being able to characterize relationships of images from different camera views plays a critical role in the performance of the person re-identification system. In this section, we address this challenge and describe the proposed attention-based feature encoder framework, as shown in Fig. 3.

First, the probe and gallery images are fed into a CNN (i.e., the Inception-V3 [60]) to generate a feature representation of each individual image. Then this feature representation is further encoded using the proposed attention-based gallery and probe encoder, taking into account the cross-view and intra-view relationship between images. The proposed feature encoder model produces highly discriminative feature representations which are encoded based on the final goal of the model (i.e., attention-based matching) and thus significantly improves the re-identification accuracy. In the following sections, we elaborate on the details of the gallery encoder and probe encoder in the proposed architecture.

3.1.1 Gallery Encoder

Let $\hat{y_i}$ denote the probe image, $\{y_j\}_{j=1}^{m}$ denote the images in the gallery set, and $f(.)$ denote the CNN embedding function. The goal of the gallery encoder is to encode the feature representation of each image in the gallery G, in the context of all the gallery images. For each feature representation in the gallery $f(y_j)$, the gallery encoder finds the feature representation of that gallery image with respect to feature representations of the rest of images in gallery (i.e., $\{f(y_g)\}_{g=1}^{m-1}, g \neq j$). In order to incorporate this information in feature representation of each gallery image, we exploit a *content*-based attention mechanism. The content-based attention has the property that the encoded feature representation will not be sensitive to the order of the images in the gallery set. In other words, the similarity information retrieved from our gallery set would not change if we randomly shuffle the images in the gallery. This is crucial for proper treatment of the gallery images.

For each gallery image $f(y_j)$ the gallery encoder function $\Psi(f(y_j), G)$ is defined as:

$$\Psi(f(y_j), G) = att(f(y_j), G, T) + f(y_j) \tag{1}$$

where $att(f(y_j), G, T)$ is an LSTM with constant input (i.e., $f(y_j)$) performing T steps of computation over the gallery images while it keeps updating its state by reading from gallery repeatedly using the attention mechanism (Eq.4 - Eq.6). In this way, the gallery encoder incorporates the contextual relationship between gallery images into the feature representation of each gallery image. The final result of encoding is the last hidden state of the LSTM after T processing steps (i.e., $att(f(y_j), G, T) = \tilde{h}_T$) plus the feature representation of the gallery image $f(y_j)$. The state of LSTM after k processing step is formulated as follows:

$$\tilde{h}_k = LSTM(f(y_j), [h_{k-1}, r_{k-1}], c_{k-1}), \tag{2}$$

$$h_k = \tilde{h}_k + f(y_j), \tag{3}$$

where in $LSTM(y, h, c)$, y being the input, h the output, and c the cell. r_{k-1} is the readout vector of the attention model and is calculated as follows:

$$r_{k-1} = \sum_{g=1}^{m-1} a(h_{k-1}, f(y_g)) f(y_g), \tag{4}$$

$$a(h_{k-1}, f(y_g)) = \frac{\exp(-d_{att}(h_{k-1}, f(y_g))}{\sum_{n=1, n\neq j}^{m-1} \exp(-d_{att}(h_{k-1}, f(y_n))}, \tag{5}$$

$$d_{att}(h_{k-1}, f(y_g)) = \|h_{k-1} - f(y_g)\|_2 \tag{6}$$

$a(h_{k-1}, f(y_g))$ in Eq. 5 is the attention weight which is calculated based on similarity (i.e., d_{att}) of h_{k-1} and the feature of each image (i.e., $f(y_g)$) in the gallery G (Eq. 6). T is the fixed number of unrolling steps of the LSTM and in order to incorporate the attention information to the LSTM, the readout attention vector r_{k-1} is concatenated with h_{k-1} in each *process* step of LSTM (Eq. 2). In fact, $[h_{k-1}, r_{k-1}]$ is the state which the LSTM evolves and k is the index which indicates how many *processing steps* are being carried to compute the state.

It is worth noting that it is possible to use $\tilde{h}_T$ as the output of the gallery encoder, but our experiments showed that it would not be as effective as our proposed scheme. By using the gallery encoder we are incorporating the view point variation relationship between images, and the gallery has the ability to modify the feature representation of each of its members based on this contextual relationship. This leads to generating rich and representative features which highly improve the re-identification accuracy. The details of the gallery encoder framework is shown in the supplementary material with a toy example.

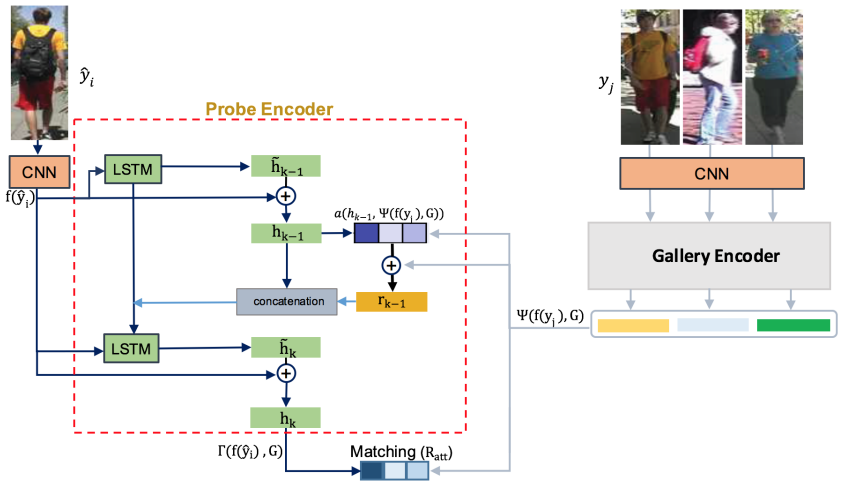

Figure 3: Details of the proposed feature encoder architecture. This is a toy example for 3 images in the gallery. Left (in red dashed box): Probe encoder. The example shows one processing step of LSTM (i.e., there exist only one LSTM in probe encoder). Right: Gallery encoder. Detail of the gallery encoder is shown by a toy example in the supplementary material of this paper. Best viewed in color.

3.1.2 Probe Encoder

The probe encoder incorporates the contextual cross-view information into feature representation of each probe image. This is performed for each probe image $\hat{y}_i$, by applying the attention mechanism over the encoded feature representation of the gallery images. For a given probe image $\hat{y}_i$ and images in the gallery $\{y_j\}_{j=1}^m$, the formulation of the probe encoder follows the same attention-based framework as in gallery encoder, with minor differences.

$$\Gamma(f(\hat{y}_i), G) = att(f(\hat{y}_i), G, T) + f(\hat{y}_i) \tag{7}$$

$$\tilde{h}_k = LSTM(f(\hat{y}_i), [h_{k-1}, r_{k-1}], c_{k-1}), \tag{8}$$

$$h_k = \tilde{h}_k + f(\hat{y}_i) \tag{9}$$

$$r_{k-1} = \sum_{j=1}^{m} a(h_{k-1}, \Psi(f(y_j), G))\Psi(f(y_j), G), \tag{10}$$

$$a(h_{k-1}, \Psi(f(y_j), G)) = \frac{\exp(-d_{att}(h_{k-1}, \Psi(f(y_j), G))}{\sum_{n=1}^{m} \exp(-d_{att}(h_{k-1}, \Psi(f(y_n), G))}, \tag{11}$$

$$d_{att}(h_{k-1}, \Psi(f(y_j), G)) = \|h_{k-1} - \Psi(f(y_j), G)\|_2 \tag{12}$$

Based on these equations we can observe that different from the gallery encoder, in the probe encoder, the input of the LSTM is the feature representation of the probe image (Eq. 8). Also, the attention is measured between the probe feature representation and all the *encoded* gallery images (i.e., $\Psi(f(y_j), G)$)). Fig. 3 shows the details of the probe encoder framework and its relationship with the rest of the model. Thus, thanks to the attention-based encoder, the encoded probe representation $\Gamma(f(\hat{y}_i), G)$ now has the information about its similarity with all the images in the gallery from different views. In this way, we are allowing the network to change its encoding of the probe image as a function of the gallery set images. This valuable information makes the encoded feature representation highly effective for the re-identification task. It is worth noting that by performing the attention mechanism over $\Psi(f(y_j), G)$, we are also incorporating the information about the relationship between the images of different people in the gallery which is encoded in $\Psi(f(y_j), G)$.

3.2. Attention-based Matching

Given $\hat{y}_i$ as the i-th probe image and $\{y_j\}_{j=1}^m$ as a set of m images in the gallery G, the goal is to match $\hat{y}_i$ with one of the $\{y_1, \ldots, y_m\}$ images. The proposed attention mechanism in our model fully specifies the re-identification task, and given the probe image $\hat{y}_i$, it generates a score for each image in the gallery set G as follows:

$$\alpha_{ij} = \|\Gamma(f(\hat{y}_i), G) - \Psi(f(y_j), G)\|_2, \tag{13}$$

where $\Gamma(f(\hat{y}_i), G)$ and $\Psi(f(y_j), G)$ are the encoded features of the $\hat{y}_i$ and the gallery image y_j, respectively (described in Sec. 3.1). After calculating the α_{ij} for each $\hat{y}_i$ and all the $\{y_j\}_{j=1}^m$ images in the gallery set, the re-identification score s_{ij} is calculated by the R_{att} function which is a softmax over α_{ij}, as:

$$s_{ij} = R_{att}(\alpha_{ij}) = \frac{\exp(-\alpha_{ij})}{\sum_{n=1}^{m} \exp(-\alpha_{in})} \tag{14}$$

The score s_{ij} is interpreted as the probability of the j-th image in the gallery set G to be a match for the probe im-

age $\hat{y}_i$. In other words, we can describe this as an associative memory case where, given an input, we "attend" to the corresponding example in the memory with highest re-identification probability and retrieve the identity of it.

3.3. Meta-learning for Few-shot Re-identification

Early approaches to meta-learning date back to the late 1980s and early 1990s, including the work by Yoshua and Samy Bengio [7]. Meta-learning has been recently used in several interesting applications such as hyper-parameter [38] and neural network optimization [28, 12], finding an optimum network architectures [40], and fast reinforcement learning [67]. The main principle of meta-learning is the ability to learn at two levels. One level is learning within each task presented using a *learner*. This learning is guided by the *meta-learner* (i.e., second level) that accumulates knowledge about the similarities and differences across different tasks. This differs from standard machine learning techniques, which involve training on a single task and testing on held-out examples from that same task.

In this paper, we use meta-learning for designing the proposed few-shot re-identification model. In our model, the attention-based matching function (discussed in Sec. 3.2) plays the role of the *learner* and the meta-learner is the neural network (i.e., the CNN and LSTM) which is trained using stochastic gradient descent. In fact, our meta-learning based model involves combining neural networks with a non-parametric structure and learning a metric space in which learning is particularly efficient. Note that the non-parametric structure here is referring to the use of gallery set as memory and using attention mechanism to retrieve information from it, as opposed to having to try and encode everything that has been seen solely in trained weights [64].

Given a probe image $\hat{y}_i$ and the gallery G, we can use the attention-based matching function R_{att} in Eq. 14, to specify a probability distribution over the gallery set identities. We formulate this mapping to be of the form $P_\theta(.|\hat{y}_i, G)$, where P is parameterized by the meta-learner (i.e., the CNN and LSTM based feature encoding network) and involves extraction of information learned *across* all the matching tasks. In this way, meta-learner can guide the learner for the new matching tasks in testing. In other words, given a new gallery set in testing, our model simply uses the parametric neural network defined by P to make predictions about the identity of each probe image. Following the meta-learning training framework in [46] and [64], our meta-learning re-identification system is trained by being exposed to several matching tasks and is then tested in its ability to learn new matching tasks (i.e., recognizing new identities in the gallery set during the testing stage) (Fig. 4). Importantly, our model does not need any fine tuning on the identities it has never seen due to its non-parametric nature. More details about the training and testing of our proposed system will be discussed next.

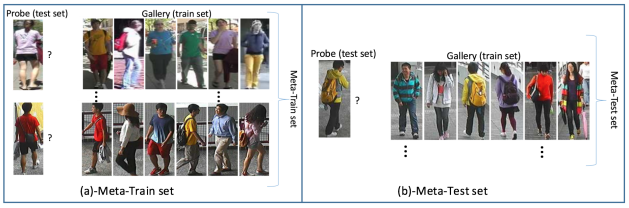

Figure 4: Example of meta-learning setup. (a) Meta-Train set: consists of multiple episodes. Each episode has one probe set and one gallery set. (b) Meta-Test set is defined in the same way as the Meta-Train set, but with a different set of probe-gallery identities not present in any of the episodes of Meta-Train set. This examples shows a 1-shot, 6-way re-identification task.

3.3.1 Training and Testing

Our meta-learning based model uses identical training and testing stages using meta-train and meta-test sets of the data. Fig. 4 shows an example of the meta-training and meta-testing sets in one-shot learning setting. In the meta-training stage, we are interested in training a neural network (the meta-learner) that can take as input one of its gallery sets G and produce a matching function (the learner) that achieves high average identification performance on its corresponding probes. The proposed model utilizes sample mini-batches called *episodes* during training where each episode is designed to mimic the testing stage by sampling labels as well as the training images. We define a task T as a uniform distribution over labels having unique classes and a few examples per class. In each episode, we sample from T a unique label set L, and then we sample from L to create each probe batch B_{probe} and the gallery set G. Note that both B_{probe} and G are the labeled examples of several identities. During training the model learns to predict the labels (identification) in the batch B_{probe} conditioned on the data in the gallery G by maximizing $\sum_{(y,z)\in B_{probe}} \log P_\theta(z|y, G)$.

The (y, z) are the image-label pairs and θ are the parameters of the feature encoding neural networks. This loss is simple and differentiable and the parameters can be optimized in an end-to-end manner. It is worth noting that in the meta-testing stage the images in the gallery set are images from new identities that the network has not been trained on before. Optimizing a model which is totally aligned with the re-identification task at the test time leads to major improvement in generalization and accuracy compared to the state-of-the-art.

4. Experiments

4.1. Datasets

There are several benchmark datasets for evaluating different person re-identification algorithms. In this paper, we use PRID2011 [23], CUHK01 [29], CUHK03 [30], and

Market1501 [84]. CUHK01, CUHK03, and Market1501 are of the largest benchmark datasets suitable for training the deep convolutional network. The following is a brief description of these four datasets:

PRID2011 dataset consists of $7,413$ images recorded by two static surveillance cameras. Camera views A and B contain 385 and 749 persons, respectively, with 200 persons appearing in both views. In our experiments on this dataset, 100 persons are used for testing.

CUHK01 dataset contains 971 persons captured from two camera views in a campus environment. Each person has four images with two from each camera. Based on standard protocol in [30], we use 871 persons for training, and the rest for testing.

CUHK03 dataset contains $13,164$ images of $1,360$ identities. Each identity has 4.8 images on average for each view. In our experiments, we use the cropped person images in this dataset. Following the conventional experimental setup (e.g., [5]), $1,160$ persons are used for training and 100 persons are used for testing.

Market1501 dataset contains $32,688$ bounding boxes of $1,501$ identities, most of which are cropped by an automatic pedestrian detector. Each person is captured by 2 to 6 cameras and has 3.6 images on average at each viewpoint. In our experiments, 750 identities are used for training and the remaining 751 for testing.

4.2. Implementation and Evaluation Protocol

We implement our model using the TensorFlow [18] deep learning framework on an Intel Xeon CPU and two NVIDIA TITAN X GPU. We use the Inception-V3 [60] structure, pre-trained on Image Net data [49] as the input embedding function (i.e., $f(.)$). Inception-V3 is a 48-layer deep convolutional architecture and since it employs global average pooling instead of fully-connected layer, it can operate on arbitrary input image sizes. The output of the last Inception block is aggregated via global average pooling to produce a 2048-D feature vector.

We use Adam [26] for optimizing the whole model. The initial learning rate is set as 0.001, and is gradually reduced after every $2,000$ iterations. We trained the model for 80 epochs. We fix the evaluation protocol across all experiments. We define the problem as an N_s-shot, N_c-way identification problem setting where a meta-learner trains on many related but small training sets (i.e., episodes) of N_s examples for each of N_c identities. For each of the four datasets, we first split the list of all identities in the data into disjoint sets and assign them to each meta-set of meta-training, and meta-testing. To generate each instance of an N_s-shot, N_c-class task episode (i.e., B_{probe}, G), we first sample N_c classes (i.e., identities) from the label set corresponding to the meta-set. We then sample N_s examples from each of those classes and these examples together compose the gallery set G. Then, an additional batch of the rest of the examples are sampled to yield a probe set (B_{probe}). When training the meta-learner, we iterate by sampling these episodes (on non-overlapping classes) repeatedly.

We used batch size of 32 in training the 1-shot setting and batch size of 16 in training the 5-shot setting. The number of processing steps for the LSTM-based attention in both encoders is $T = 100$. We crop a center region of the person's image with a small random perturbation from each image to augment the data (for a few identities in some datasets) when there are less than 6 examples per ID. We report rank1, rank5, rank10 and rank20 accuracies of cumulative match curve (CMC) on four datasets to evaluate the proposed re-identification model performance. All the following results are based on **Single Query (SQ)** and **without re-ranking** in testing stage. It is worth noting that for all the experiments 100-way setting is used for the testing episodes and for the Market 1501 data with 751 testing samples, the testing results are averaged across episodes.

4.3. Evaluation of Individual Components

The proposed ARM has three novel components, including the attention-based matching framework, the attention-based probe encoder, and the attention-based gallery encoder. To illustrate how each component contributes to the performance improvement of the whole system, we implement the following five variants of our model:

- The full version of the proposed Attention-based Re-identification Meta-learning Model with all the components, denoted as ARM.
- The attention-matching with only the probe encoder (i.e., no gallery encoder), denoted as ARM1.
- The attention-matching with only the gallery encoder (i.e., no probe encoder), denoted by ARM2.
- The attention-matching without any encoder (i.e., no probe and gallery encoder), denoted by ARM3.
- The matching model (without meta-learning and encoder but in 5-shot setting) using the same branches for probe and gallery images (Siamese net.), denoted by Baseline.

For this part of the experiment, we compare the rank1 performance for all the above 4 scenarios on PRID2011, CUHK01, CUHK03, and Market1501 datasets (in 5-shot learning). The results in Table 1 illustrate that the full version of the attention-based re-identification model (ARM) outperforms all the other variants, which shows the importance and effectiveness of all the components of the model. As we can see in Table 1, for all 4 datasets, removing either the gallery encoder $\Psi(f(y_j), G)$ (i.e., ARM1) or the probe

Table 1: Comparison of rank1 performance (%) of different variations of the proposed ARM model (5-shot learning) on 4 datasets.

	PRID2011	CUHK01	CUHK03	Market1501
ARM	**87.94**	**95.42**	**91.98**	**90.10**
ARM1	86.73	89.62	82.88	86.25
ARM2	70.01	87.00	75.81	76.22
ARM3	68.63	83.29	72.08	71.63
Baseline	40.12	61.20	50.23	44.98

encoder $\Gamma(f(\hat{y}_i), G)$ (i.e., ARM2), causes the accuracy to drop. However, Table 1 shows that removing the probe encoder (ARM2) leads to higher accuracy drop than removing the gallery encoder (ARM1). This is due to the fact that even by removing the gallery encoder, we are still able to incorporate some of the cross-view information by attending to the features of gallery images, even though we are losing the contextual information about intrinsic relationship between the gallery images. On the other hand, by removing the probe encoder, the gallery encoder will not be able to modify the feature representation of the probe image and we lose all the critical cross-view information for recognizing people across cameras. Furthermore, since the structure of the probe encoder reflects the final goal of the model, it produces representative features which are specially modified for the re-identification task. By removing the whole encoder (i.e., ARM3), we are ignoring the intra-view dependency between images in the gallery set as well as the relationship between the probe and gallery set images. By using the proposed encoder framework, our model adapts to differences in the camera views (e.g., background, lighting differences, etc.), and therefore the differences and similarities of the feature representation of different pedestrians would be the focus of the model for performing the matching process. Moreover, the memory in the LSTM module, used in the encoder, helps the meta-learning framework to better guide the learner in different tasks. Finally removing the encoders and training the model as a Siamese network in Baseline showed drastic drop of accuracy in all ranks.

4.3.1 Number of Shots, Distance Metric and Computational Complexity

As it was discussed before, we construct episodes, by choosing N_c identities and N_s examples per identity (i.e., N_s-shot, N_c-way) during training. Experiments in all 4 datasets show that the rank1 accuracy increases when using 5-shot (i.e., $N_s = 5$) compared to one-shot case (Table 2), since by using multiple examples of each identity the intra-class variation is taken into account. However, even by using one example per identity, our model still accurately performs the re-identification task. Moreover, our experiments show that using Euclidean distance as the measure of similarity in our attention models (e.g., Eq. 6) leads to superior performance compared to commonly used cosine distance. Table 2 shows the rank1 testing accuracy of the proposed ARM method on PRID11, CUHK01, CUHK03, and Market1501 datasets in 100-way, 1-shot and 5-shot (i.e., $N_s = 1, 5$ and $N_c = 100$) scenarios and by using cosine vs. Euclidean distance.

Table 2: Rank1 accuracy for 1-shot and 5-shot identification using the proposed ARM.

N_s-shot / Dist. metric	PRID2011	CUHK01	CUHK03	Market1501
1-shot/Cosine	61.99	83.50	79.74	72.93
1-shot/Euclid.	65.32	83.56	80.08	76.99
5-shot/Cosine	81.90	94.00	87.22	85.91
5-shot/Euclid.	**87.94**	**95.42**	**91.98**	**90.10**

We also study the computational cost of using more examples in the gallery set during training. By using the hardware and software framework explained in Sec. 4.2, our experiments show that the ARM model on average needs 269.93 seconds per epoch on 1-shot setting and 860.80 seconds per epoch on 5-shot setting for training on Market1501 dataset. The test time for each image is around 2.1 seconds. As the gallery set size grows in size, the computation for each gradient update becomes more expensive and solving this problem is our future research direction.

4.4. Comparison with State-of-the-Art

Table 3: Rank1 accuracy (%) comparison of the proposed ARM method to the state-of-the-art on PRID2011, CUHK01, CUHK03 and Market1501 dataset in 1-shot and 5-shot setting.

Method	PRID2011	CUHK01	CUHK03	Market1501
BraidNet-CS + SRL[69]	-	93.04	88.18	83.70
ImprovedDL [1]	-	65.00	54.74	-
EDM [54]	-	86.59	61.32	-
MTDnet[11]	-	78.50	74.68	-
DNNIM [58]	-	81.23	72.43	-
DCSL [79]	-	89.60	80.20	-
LS-CNN [62]	-	-	57.30	61.60
SCSP [9]	-	-	-	51.90
Spindle [80]	67.00	-	88.50	76.09
SIR-CIR [66]	-	71.80	52.20	-
KISSME[27]	15.00	29.40	14.27	-
PaMM (M) [14]	45.00	-	-	-
DNS [78]	-	-	-	55.4
MuDeep[44]	65.00	79.01	76.09	-
DeepM[76]	17.90	-	-	-
GS-CNN [61]	-	-	61.08	65.88
DGDrop [71]	64.00	-	75.03	59.53
DCIA [20]	39.00	-	-	-
PDC [55]	-	-	88.70	84.14
SSM [5]	72.98	-	76.60	82.21
PrtAlign [81]	-	88.50	81.60	81.00
PSE [52]	-	-	-	87.70
P2S [86]	-	77.34	-	70.72
TriNet [22]	-	-	-	84.92
JLML [31]	-	87.00	83.20	85.10
AWTL (2-stream) [47]	-	-	-	89.46
AACN [74]	-	-	90.58	88.70
ARM, 1-shot (ours)	**65.32**	**83.56**	**80.08**	**76.99**
ARM, 5-shot (ours)	**87.94**	**95.42**	**91.98**	**90.10**

Table 4 shows the performance comparison of the proposed ARM approach with state-of-the-art in 1-shot, 5-shot setting on 4 different datasets. Some of the reported scores for other works are taken from [69], following the same ex-

perimental setting. It is worth noting that unlike the reported state-of-the-art re-identification methods, even though our proposed method uses only 1 example per ID, it is still able to achieve better or comparable accuracy in most cases (unlike the other works that need many samples per identity for training). Also, in 5-shot setting ARM outperforms all the other methods. In Table 4, the best one-shot re-identification results are marked with red color and the best results using other learning methods (including our 5-shot method) are shown in blue. We also observe that ARM outperforms the method in [13] and [22] which are based on using the improved triplet loss metric. Furthermore, ARM performs better than DCIA [20] which is a related work to ours that considers the relationship between gallery images in a ranking framework. Due to lack of space in the paper, we only compare rank1 accuracy for all the methods and datasets and the complete results for all ranks are included in the supplementary material.

5. Conclusion

This study represents the first attempt for solving the person re-identification problem in a few-shot learning framework by exploiting the concept of meta-learning. All the three main components of our model were designed based on the attention mechanism and the importance of each component was demonstrated using extensive experiments. Furthermore, we provided analysis showing that some design choices can yield substantial improvements compared to existing models. Our future work is to extend the proposed framework to *video memory machine* for (end-to-end) few-shot video-based person re-identification.

References

[1] E. Ahmed, M. Jones, and T. K. Marks. An improved deep learning architecture for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3908–3916, 2015. 1, 3, 8, 14

[2] J. Ba, G. E. Hinton, V. Mnih, J. Z. Leibo, and C. Ionescu. Using fast weights to attend to the recent past. In *Advances In Neural Information Processing Systems*, pages 4331–4339, 2016. 3

[3] J. Ba, V. Mnih, and K. Kavukcuoglu. Multiple object recognition with visual attention. *arXiv preprint arXiv:1412.7755*, 2014. 3

[4] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014. 3

[5] S. Bai, X. Bai, and Q. Tian. Scalable person re-identification on supervised smoothed manifold. *arXiv preprint arXiv:1703.08359*, 2017. 7, 8

[6] S. Bak and P. Carr. One-shot metric learning for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 3

[7] Y. Bengio, S. Bengio, and J. Cloutier. *Learning a synaptic learning rule*. Université de Montréal, Département d'informatique et de recherche opérationnelle, 1990. 6

[8] J. Bromley, J. W. Bentz, L. Bottou, I. Guyon, Y. LeCun, C. Moore, E. Säckinger, and R. Shah. Signature verification using a siamese time delay neural network. *International Journal of Pattern Recognition and Artificial Intelligence*, 7(04):669–688, 1993. 3

[9] D. Chen, Z. Yuan, B. Chen, and N. Zheng. Similarity learning with spatial constraints for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1268–1277, 2016. 8, 14

[10] W. Chen, X. Chen, J. Zhang, and K. Huang. Beyond triplet loss: a deep quadruplet network for person re-identification. *arXiv preprint arXiv:1704.01719*, 2017. 2, 3, 14

[11] W. Chen, X. Chen, J. Zhang, and K. Huang. A multi-task deep network for person re-identification. In *AAAI*, pages 3988–3994, 2017. 8

[12] Y. Chen, M. W. Hoffman, S. G. Colmenarejo, M. Denil, T. P. Lillicrap, M. Botvinick, and N. de Freitas. Learning to learn without gradient descent by gradient descent. *arXiv preprint arXiv:1611.03824*, 2016. 6

[13] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1335–1344, 2016. 1, 2, 3, 9, 13

[14] Y.-J. Cho and K.-J. Yoon. Improving person re-identification via pose-aware multi-shot matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1354–1362, 2016. 8, 13

[15] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon. Information-theoretic metric learning. In *Proceedings of the 24th international conference on Machine learning*, pages 209–216. ACM, 2007. 13

[16] S. Ding, L. Lin, G. Wang, and H. Chao. Deep feature learning with relative distance comparison for person re-identification. *Pattern Recognition*, 48(10):2993–3003, 2015. 2

[17] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2625–2634, 2015. 3

[18] M. A. et al. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org. 7

[19] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. Person re-identification by symmetry-driven accumulation of local features. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2360–2367. IEEE, 2010. 14

[20] J. Garcia, N. Martinel, C. Micheloni, and A. Gardel. Person re-identification ranking optimisation by discriminant context information analysis. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1305–1313, 2015. 3, 8, 9

[21] M. Guillaumin, J. Verbeek, and C. Schmid. Is that you? metric learning approaches for face identification. In *Computer Vision, 2009 IEEE 12th international conference on*, pages 498–505. IEEE, 2009. 14

[22] A. Hermans, L. Beyer, and B. Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017. 8, 9

[23] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof. Person re-identification by descriptive and discriminative classification. In *Scandinavian conference on Image analysis*, pages 91–102. Springer, 2011. 6

[24] M. Hirzer, P. M. Roth, and H. Bischof. Person re-identification by efficient impostor-based metric learning. In *Advanced Video and Signal-Based Surveillance (AVSS), 2012 IEEE Ninth International Conference on*, pages 203–208. IEEE, 2012. 13, 14

[25] Ł. Kaiser, O. Nachum, A. Roy, and S. Bengio. Learning to remember rare events. *ICLR-arXiv:1703.03129*, 2017. 3

[26] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 7

[27] M. Köstinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof. Large scale metric learning from equivalence constraints. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2288–2295. IEEE, 2012. 8, 13, 14

[28] D. Li, X. Chen, Z. Zhang, and K. Huang. Learning deep context-aware features over body and latent parts for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 384–393, 2017. 3, 6

[29] W. Li and X. Wang. Locally aligned feature transforms across views. pages 3594–3601, 2013. 6

[30] W. Li, R. Zhao, T. Xiao, and X. Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 152–159, 2014. 3, 6, 7, 14

[31] W. Li, X. Zhu, and S. Gong. Person re-identification by deep joint learning of multi-loss classification. *arXiv preprint arXiv:1705.04724*, 2017. 8

[32] Z. Li, S. Chang, F. Liang, T. S. Huang, L. Cao, and J. R. Smith. Learning locally-adaptive decision functions for person verification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3610–3617, 2013. 3

[33] S. Liao, Y. Hu, X. Zhu, and S. Z. Li. Person re-identification by local maximal occurrence representation and metric learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2197–2206, 2015. 3

[34] S. Liao, G. Zhao, V. Kellokumpu, M. Pietikäinen, and S. Z. Li. Modeling pixel process with scale invariant local patterns for background subtraction in complex scenes. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 1301–1306. IEEE, 2010. 3

[35] H. Liu, J. Feng, M. Qi, J. Jiang, and S. Yan. End-to-end comparative attention networks for person re-identification. *arXiv preprint arXiv:1606.04404*, 2016. 3

[36] H. Liu, M. Qi, and J. Jiang. Kernelized relaxed margin components analysis for person re-identification. *IEEE Signal Processing Letters*, 22(7):910–914, 2015. 14

[37] Z. Liu, D. Wang, and H. Lu. Stepwise metric promotion for unsupervised video person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2429–2438, 2017. 2

[38] D. Maclaurin, D. Duvenaud, and R. Adams. Gradient-based hyperparameter optimization through reversible learning. In *International Conference on Machine Learning*, pages 2113–2122, 2015. 6

[39] N. Martinel, A. Das, C. Micheloni, and A. K. Roy-Chowdhury. Temporal model adaptation for person re-identification. In *European Conference on Computer Vision*, pages 858–877. Springer, 2016. 13

[40] R. Negrinho and G. Gordon. Deeparchitect: Automatically designing and training deep architectures. *arXiv preprint arXiv:1704.08792*, 2017. 6

[41] T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on pattern analysis and machine intelligence*, 24(7):971–987, 2002. 3

[42] S. Pedagadi, J. Orwell, S. Velastin, and B. Boghossian. Local fisher discriminant analysis for pedestrian re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3318–3325, 2013. 3

[43] W. Pei, D. M. Tax, and L. van der Maaten. Modeling time series similarity with siamese recurrent networks. *arXiv preprint arXiv:1603.04713*, 2016. 3

[44] X. Qian, Y. Fu, Y.-G. Jiang, T. Xiang, and X. Xue. Multi-scale deep learning architectures for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5399–5408, 2017. 8, 13, 14

[45] A. Rahimpour, L. Liu, A. Taalimi, Y. Song, and H. Qi. Person re-identification using visual attention. In *IEEE International Conference on Image processing*. IEEE, 2017. 3

[46] S. Ravi and H. Larochelle. Optimization as a model for few-shot learning. ICLR 2017. 6

[47] E. Ristani and C. Tomasi. Features for multi-target multi-camera tracking and re-identification. *arXiv preprint arXiv:1803.10859*, 2018. 8

[48] P. M. Roth, M. Hirzer, M. Köstinger, C. Beleznai, and H. Bischof. Mahalanobis distance learning for person re-identification. In *Person Re-Identification*, pages 247–267. Springer, 2014. 13

[49] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. 7

[50] A. Santoro, S. Bartunov, M. Botvinick, D. Wierstra, and T. Lillicrap. Meta-learning with memory-augmented neural networks. In *International conference on machine learning*, pages 1842–1850, 2016. 3

[51] A. Santoro, S. Bartunov, M. Botvinick, D. Wierstra, and T. Lillicrap. One-shot learning with memory-augmented neural networks. *arXiv preprint arXiv:1605.06065*, 2016. 3

[52] M. Saquib Sarfraz, A. Schumann, A. Eberle, and R. Stiefelhagen. A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 420–429, 2018. 8

[53] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015. 2

[54] 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 *European Conference on Computer Vision*, pages 732–748. Springer, 2016. 8

[55] C. Su, J. Li, S. Zhang, J. Xing, W. Gao, and Q. Tian. Pose-driven deep convolutional model for person re-identification. *ICCV*, 2017. 8, 14

[56] C. Su, F. Yang, S. Zhang, Q. Tian, L. S. Davis, and W. Gao. Multi-task learning with low rank attribute embedding for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3739–3747, 2015. 3

[57] C. Su, S. Zhang, J. Xing, W. Gao, and Q. Tian. Deep attributes driven multi-camera person re-identification. In *European Conference on Computer Vision*, pages 475–491. Springer, 2016. 2

[58] A. Subramaniam, M. Chatterjee, and A. Mittal. Deep neural networks with inexact matching for person re-identification. In *Advances in Neural Information Processing Systems*, pages 2667–2675, 2016. 8

[59] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014. 3

[60] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826, 2016. 4, 7

[61] R. R. Varior, M. Haloi, and G. Wang. Gated siamese convolutional neural network architecture for human re-identification. In *European Conference on Computer Vision*, pages 791–808. Springer, 2016. 3, 8, 14

[62] R. R. Varior, B. Shuai, J. Lu, D. Xu, and G. Wang. A siamese long short-term memory architecture for human re-identification. In *European Conference on Computer Vision*, pages 135–153. Springer, 2016. 3, 8, 14

[63] R. R. Varior, G. Wang, J. Lu, and T. Liu. Learning invariant color features for person re-identification. In *IEEE Transaction on Image processing*. IEEE, 2016. 1, 3

[64] O. Vinyals, C. Blundell, T. Lillicrap, D. Wierstra, et al. Matching networks for one shot learning. In *Advances in Neural Information Processing Systems*, pages 3630–3638, 2016. 3, 6

[65] O. Vinyals, M. Fortunato, and N. Jaitly. Pointer networks. In *Advances in Neural Information Processing Systems*, pages 2692–2700, 2015. 3

[66] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2, 3, 8, 14

[67] J. X. Wang, Z. Kurth-Nelson, D. Tirumala, H. Soyer, J. Z. Leibo, R. Munos, C. Blundell, D. Kumaran, and M. Botvinick. Learning to reinforcement learn. *arXiv preprint arXiv:1611.05763*, 2016. 6

[68] W. Wang, A. Taalimi, K. Duan, R. Guo, and H. Qi. Learning patch-dependent kernel forest for person re-identification. In *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*, pages 1–9. IEEE, 2016. 14

[69] Y. Wang, Z. Chen, F. Wu, and G. Wang. Person re-identification with cascaded pairwise convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1470–1478, 2018. 8

[70] J. Weston, S. Chopra, and A. Bordes. Memory networks. *arXiv preprint arXiv:1410.3916*, 2014. 3

[71] T. Xiao, H. Li, W. Ouyang, and X. Wang. Learning deep feature representations with domain guided dropout for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 3, 8, 14

[72] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang. Joint detection and identification feature learning for person search. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3376–3385. IEEE, 2017. 3

[73] F. Xiong, M. Gou, O. Camps, and M. Sznaier. Person re-identification using kernel-based metric learning methods. In *European Conference on Computer Vision*, pages 1–16. Springer, 2014. 3

[74] J. Xu, R. Zhao, F. Zhu, H. Wang, and W. Ouyang. Attention-aware compositional network for person re-identification. *arXiv preprint arXiv:1805.03344*, 2018. 8

[75] K. Xu, J. Ba, R. Kiros, K. Cho, A. C. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, volume 14, pages 77–81, 2015. 3

[76] D. Yi, Z. Lei, S. Liao, S. Z. Li, et al. Deep metric learning for person re-identification. In *ICPR*, volume 2014, pages 34–39, 2014. 1, 8, 13, 14

[77] J. Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici. Beyond short snippets: Deep networks for video classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4694–4702, 2015. 3

[78] L. Zhang, T. Xiang, and S. Gong. Learning a discriminative null space for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1, 3, 8, 13, 14

[79] Y. Zhang, X. Li, L. Zhao, and Z. Zhang. Semantics-aware deep correspondence structure learning for robust person re-identification. In *IJCAI*, pages 3545–3551, 2016. 8

[80] H. Zhao, M. Tian, S. Sun, J. Shao, J. Yan, S. Yi, X. Wang, and X. Tang. Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1077–1085, 2017. 3, 8, 13, 14

[81] L. Zhao, X. Li, Y. Zhuang, and J. Wang. Deeply-learned part-aligned representations for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3219–3228, 2017. 8, 14

[82] R. Zhao, W. Ouyang, and X. Wang. Person re-identification by salience matching. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2528–2535, 2013. 3

[83] R. Zhao, W. Ouyang, and X. Wang. Unsupervised salience learning for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3586–3593, 2013. 2, 14

[84] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1116–1124, 2015. 3, 6

[85] L. Zheng, H. Zhang, S. Sun, M. Chandraker, and Q. Tian. Person re-identification in the wild. *arXiv preprint arXiv:1604.02531*, 2016. 1, 3

[86] S. Zhou, J. Wang, J. Wang, Y. Gong, and N. Zheng. Point to set similarity based deep feature learning for person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 3, 8, 14

[87] Z. Zhou, Y. Huang, W. Wang, L. Wang, and T. Tan. See the forest for the trees: Joint spatial and temporal recurrent neural networks for video-based person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4747–4756, 2017. 3

A. Appendix.

A.1. Toy Example: Attention-Based Gallery Encoder

This supplementary material comprises a toy example to demonstrate the details of the gallery encoder and additional results and comparisons that have not been explained in the paper due to page limit.

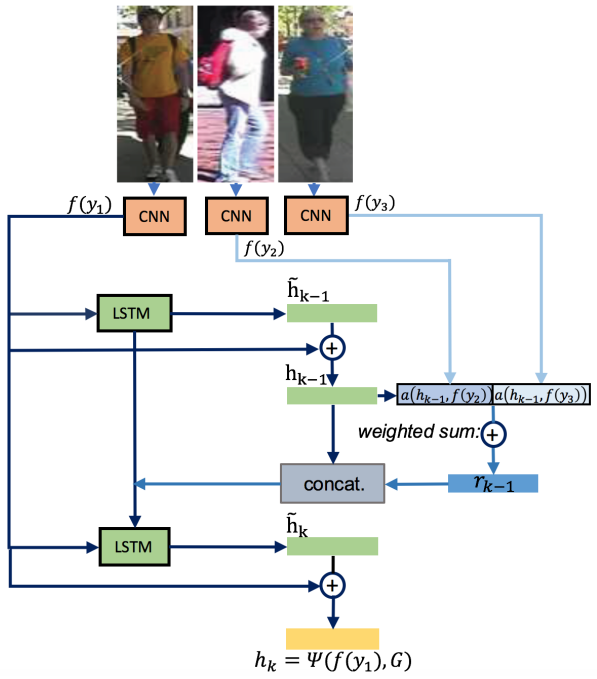

Figure 5: Gallery encoder architecture. The encoder exploits the LSTM-based attention model to incorporate the information about the relationship between the gallery images, into feature representation of each gallery image. This figure shows the last step of the encoding (i.e., k-th).

Fig. 5 shows a toy example to better explain the internal structure of the gallery encoder. In this example, the goal is to calculate the feature encoding of the first gallery image (y_1). After deriving the feature representation of all the gallery images using a CNN (i.e., $f(y_j)$), the feature representation of y_1 goes through the LSTM and the resulting hidden state of the LSTM gets processed for a fixed number of steps (e.g., k). At each step, the LSTM also attends over the feature representation of images in the gallery set. Fig. 5 shows the last processing step of the encoder (i.e., k-th). In fact, in each step, $a(h_{k-1}, f(y_2))$ and $a(h_{k-1}, f(y_3))$ are the attention weights which are calculated based on the similarity (i.e., d_{att}) of the hidden state of the LSTM (i.e., h_{k-1}) and the feature of each image (i.e., $f(y_2)$, $f(y_3)$) in the gallery G (Eq. 5). r_{k-1} is the weighted average of $f(y_2)$ and $f(y_3)$ and is concatenated with the hidden state of the LSTM in each step (Eq. 4, Eq. 3).

Hence, in each step, the hidden state of the LSTM evolves based on the relationship between gallery images and at the end (e.g., after k steps), it will contain "deep" contextual information about this relationships. The result of encoding after k steps is the last hidden state of the LSTM plus the feature representation of the image $(\tilde{h}_k + f(y_1))$.

The encoded feature of y_1 is shown as a yellow block in Fig. 5 here and in Fig. 3 in the paper. This process is performed the same way for all the other gallery images. By using the gallery encoder we are incorporating the intra-view and cross-view relationship between gallery images into feature representation of each image, and the gallery has the ability to modify the feature representation of each of its members based on this relationship.

A.2. Additional Results: Comparison with State-of-the-art for Different Ranks

In Sec. 4.4 of the paper, rank 1 accuracy of ARM is compared with state-of-the-art on 4 different datasets. In this section, additional results are provided for different ranks. We demonstrate the results for rank 1, rank 5, rank 10 and rank 20 for each dataset separately. Experimental setting is identical to what has been described in the paper.

PRID2011:

Table 4 illustrates the comparison of the results of our proposed method (ARM) with state-of-the-art results on PRID2011 dataset. The proposed ARM method (5-shot), outperforms all these methods by a large margin in all ranks.

Table 4: Performance (%) comparison of ARM with state-of-the-art on PRID2011 dataset.

Method	Rank1	Rank5	Rank10	Rank20
TMA [39]	54.02	73.8	83.01	90.02
LMNN [24]	10.0	30.0	42.0	59.0
ITML[15]	12.0	36.0	47.0	64.0
NFST [78]	40.09	64.70	73.20	81.00
Maha[48]	16.0	41.0	51.0	64.0
Spindle [80]	67.00	89.00	89.00	92.00
MuDeep[44]	65.00	87.00	93.00	-
DeepM[76]	17.90	45.90	55.40	71.40
KISSME[27]	15.00	39.00	52.00	68.00
PaMM (M) [14]	45.00	72.00	85.00	92.50
Triplet [13]	22.00	-	47.00	57.00
ARM, 1-shot (ours)	65.32	73.46	85.98	90.99
ARM, 5-shot (ours)	**87.94**	**92.00**	**94.89**	**95.10**

CUHK01 and **CUHK03:** The additional results for CUHK01 and CUHK03 are reported in Table 5 and 6. We observe that the proposed ARM method outperforms all the traditional as well as the recent deep learning-based methods.

Market1501: As Table 7 shows, ARM achieves the best results in all ranks and mean Average Precision (mAP). In our experiments, 750 identities are used for training and the remaining 751 for testing.

Table 5: Performance (%) comparison of ARM with state-of-the-art on CUHK01 dataset.

Method	Rank1	Rank5	Rank10	Rank20
LDM[21]	26.45	57.69	72.04	84.69
KISSME [27]	29.40	57.67	72.42	86.07
SDALF [19]	9.90	41.21	56.00	66.37
eSDC [83]	22.84	43.89	57.67	69.84
Patchdep.[68]	44.00	78.50	86.70	94.00
SIR-CIR [66]	71.80	91.60	96.00	98.00
PrtAlign [81]	88.50	98.40	99.60	99.90
MuDeep [44]	79.01	97.00	98.96	-
P2S [86]	77.34	93.51	96.73	98.53
ARM, 1-shot (ours)	83.56	87.34	93.15	98.49
ARM, 5-shot (ours)	**95.42**	**97.73**	**99.66**	**99.97**

Table 6: Performance (%) comparison of ARM with state-of-the-art on CUHK03 labeled dataset.

Method	Rank1	Rank5	Rank10	Rank20
LMNN[24]	7.29	21.00	32.06	48.94
KRMCA[36]	9.23	25.73	35.09	52.96
LDM[21]	13.51	40.73	52.13	70.81
FPNN [30]	20.65	51.50	66.50	80.00
eSDC [83]	8.76	24.07	38.28	53.44
KISSME [27]	14.17	48.54	52.57	70.53
Imp-reid [1]	54.74	86.50	93.88	98.10
Deep Metric [76]	61.30	88.50	96.00	99.00
SIR-CIR [66]	52.20	85.00	92.00	97.00
MuDeep(J)[44]	76.87	96.12	98.41	-
Spindle [80]	88.50	97.80	98.60	99.20
PrtAlign [81]	85.41	97.62	99.43	99.93
Quadruplet [10]	75.53	95.15	99.16	-
ARM, 1-shot (ours)	80.08	87.60	90.37	95.37
ARM, 5-shot (ours)	**91.98**	**98.74**	**99.50**	**99.98**

Table 7: Performance (%) comparison of ARM with state-of-the-art on Market1501 dataset.

Method	Rank1	Rank5	Rank10	Rank20	mAP
KISSME [27]	44.40	63.90	72.02	-	19.02
GS-CNN [61]	65.88	-	-	-	39.55
DGDrop [71]	59.53	-	-	-	31.94
LS-CNN [62]	61.60	-	-	-	35.30
SCSP [9]	51.9	72.0	79.0	-	26.35
DNS [78]	55.40	-	-	-	35.68
Spindle [80]	76.90	91.50	94.60	96.70	-
P2S [86]	70.72	-	-	-	44.27
PrtAl [81]	81.00	92.00	94.70	-	-
PDC [55]	84.14	92.73	94.92	96.82	63.41
ARM, 1-shot (ours)	76.99	78.51	88.23	93.31	65.96
ARM, 5-shot (ours)	**90.10**	**92.91**	**96.63**	**98.33**	**78.98**